https://doi.org/10.26599/JICV.2026.9210110 Research Article

VISTA+: Mesh-based single-view novel view synthesis and CAD-based foreground composition for synthetic driving-scene generation

Song Zhang[1], Ruibo Wang[1], Jian Li[1], Wenxiu Shi[1], Huiping Zhu[1], Ruo Jia[2], Ziyu Lin[3], Haiquan Liu[4,✉]

[1] *Z-one Technology Co., Ltd., Shanghai 201101, China*

[2] *Department of Architecture and Civil Engineering, Chalmers University of Technology, Göteborg SE-41296, Sweden*

[3] *School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China*

[4] *Tsing-AI (Shanghai) Technology Co., Ltd., Shanghai 200000, China*

✉ *Corresponding author.*

E-mail address: haiquan.liu@zex-t.com

Received: August 12, 2026; Revised: September 4, 2026; Accepted: October 27, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

Just Accepted

Abstract

Synthetic data has emerged as a promising solution for alleviating data scarcity in autonomous driving, especially in cold-start and corner-case scenarios. However, existing synthetic data generation methods still face a trade-off between photorealistic fidelity and engineering scalability. This paper proposes VISTA+, a framework for synthetic driving-scene generation that combines mesh-based single-view novel view synthesis (NVS) of static backgrounds with computer-aided design (CAD)-based foreground rendering and foreground-background composition. Unlike multi-view neural reconstruction methods, VISTA+ avoids large-scale multi-view scene reconstruction, while still requiring camera calibration and source-to-target relative poses, thereby reducing engineering complexity. Specifically, we improve mesh-based background NVS by incorporating foreground inpainting, dense depth supervision, structural pretraining, and joint depth-offset prediction. Dynamic foreground objects are rendered using CAD assets under estimated environmental lighting and are then integrated with generated backgrounds through image harmonization. Experiments on KITTI and nuScenes demonstrate that VISTA+ outperforms representative single-view NVS baselines in terms of peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Furthermore, synthetic data generated by

VISTA+ consistently improves the performance of multiple downstream three-dimensional (3D) object detection models, demonstrating the downstream utility and functional realism of the generated driving scenes.

Keywords: autonomous driving; synthetic data generation; single-view novel view synthesis; textured mesh; foreground-background composition

1. Introduction

End-to-end neural networks have shown strong potential for autonomous driving (Jiang et al., 2023; Sun et al., 2025; Zhu et al., 2026), but their development requires large-scale training data, the collection and annotation of which are often costly and resource-intensive. Synthetic data generated by simulation, rendering, and data-driven generation methods (Wrenninge and Unger, 2018; Tsirikoglou et al., 2017; Lin et al., 2023; Gómez et al., 2025; He et al., 2025; Zhang et al., 2026) provides a feasible way to mitigate data scarcity. Analysis of automated-vehicle crash reports identifies maneuvering, road, and lighting conditions as safety-relevant factors (Kurse et al., 2025), motivating diverse scenarios for training and testing. Nevertheless, existing synthesis pipelines remain limited by photorealistic fidelity and implementation efficiency. Images generated by real-time rendering engines still exhibit a noticeable realism gap (Dosovitskiy et al., 2017; Shah et al., 2018), and computer-aided design (CAD)-based scene modeling is labor-intensive. Data-driven methods such as Neural Radiance Fields (NeRF) (Mildenhall et al., 2022; Lin et al., 2026a), three-dimensional Gaussian Splatting (3DGS) (Chen et al., 2024; Kerbl et al., 2023), and image and video generation models (Yang et al., 2023a; Gao et al., 2023; Lin et al., 2026b) can introduce substantial data and computation requirements. Scene-specific neural reconstruction requires posed observations and optimization, whereas generative models typically require large training corpora and considerable training resources.

VISTA (Amini et al., 2020, 2022) is a data-driven simulator that synthesizes red, green, and blue (RGB) images at novel viewpoints using an image and its depth map from the spatially closest recorded pose. By avoiding computationally expensive multi-view stereo (MVS) reconstruction, VISTA substantially reduces the complexity of engineering implementation. However, the visual fidelity of its novel view synthesis (NVS) module remains limited, motivating further improvement for high-quality synthetic data generation.

Inspired by prior work (Blender Foundation, 2024; Hu et al., 2021), we develop VISTA+, a framework for synthetic driving-scene generation that combines improved mesh-based single-view NVS of static backgrounds with CAD-based foreground rendering and foreground-background composition. To enhance photorealism, the

Just Accepted

scene is decomposed into dynamic foregrounds and static backgrounds. Background regions are restored through image inpainting and represented as textured meshes, with reconstruction further improved by pretrained depth estimation, multi-task learning, and proxy training. Foreground objects such as vehicles are rendered in Blender using CAD assets and environmental lighting. Finally, foreground and background layers are fused using image harmonization to produce visually consistent novel-view images (Fig. 2).

In practical applications, images of a road segment can be pre-collected from repeated traversals. Given a desired viewpoint and the poses of surrounding traffic participants, the closest recorded image is selected and processed by the proposed NVS module to generate the background view, while foreground objects are rendered from CAD assets in Blender. This setting is well aligned with engineering scenarios in which vehicles frequently travel along the same route, such as autonomous development vehicles, fixed-route buses, and robotaxis. It is also promising for virtual reality (VR) and augmented reality (AR) applications, especially when 360-degree panoramic cameras are available. Validation of VR bicycle simulators uses both perceived realism and real-world behavioral measures (Hammami et al., 2025), highlighting the need for application-specific validation before extending VISTA+ to such settings.

The main contributions of this study are summarized as follows:

1) We propose VISTA+, a framework for synthetic driving-scene generation that integrates mesh-based single-view background NVS, CAD-based foreground rendering, and foreground-background composition.

2) We improve mesh-based background NVS by incorporating foreground inpainting, dense depth supervision, structural pretraining, and joint depth-offset prediction.

3) VISTA+ achieves superior NVS quality compared with representative single-view NVS baselines and improves downstream 3D object detection performance when used for synthetic data augmentation.

2. Related work

2.1. Single-image NVS

Single-image NVS aims to infer a 3D representation from a single input image and render novel viewpoints using monocular feed-forward reconstruction networks (Tucker and Snavely, 2020; Wiles et al., 2020; Hu et al., 2021). Representative methods use different 3D representations: Tucker and Snavely (2020) predict multi-plane images (MPIs), Li et al. (2021) extend MPI to continuous depth, Tulsiani et al. (2018) infer layered depth images, and Tulsiani et al. (2017) use a volumetric representation supervised by differentiable ray consistency. Other methods adopted neural representations, such as NeRF (Mildenhall et al., 2022), as in Wimbauer et al. (2023).

Just Accepted

SynSin (Wiles et al., 2020) combined monocular depth prediction with a rendering network to refine novel views, while Flash3D (Szymanowicz et al., 2025) integrated pretrained depth estimation with 3DGS fields to improve reconstruction fidelity. Worldsheet (Hu et al., 2021) represents scenes as textured meshes, which are particularly suitable for smooth background structures with continuous topology. Although Klinghoffer et al. (2023) used Worldsheet to enhance viewpoint robustness in autonomous-driving perception, artifacts caused by dynamic foreground objects remain a major limitation. Liu et al. (2023) combined a scaffold mesh with a learned neural texture field and shader, suggesting a possible direction for improving the appearance of mesh-based rendering. Building on the Worldsheet paradigm, this work uses generative inpainting to remove dynamic, non-smooth foreground objects before mesh reconstruction, thereby better satisfying the smooth sheet-like geometry assumption. In addition, improved depth supervision and proxy tasks are introduced to enhance reconstruction accuracy.

2.2. Foreground-background image composition

Foreground-background composition has been widely studied for generating visually plausible composite images. Niu et al. (2021) provided a comprehensive survey of deep image composition, with particular attention to image harmonization. In autonomous-driving applications, decomposing foreground and background generation has also proven effective. Abu Alhaija et al. (2018) used real photographic backgrounds to improve visual fidelity, while GeoSim (Chen et al., 2021) employed neural reconstruction to synthesize foreground objects and insert them into geometry-aware background images. Ost et al. (2021) used separate neural rendering models for foreground and background components, and OmniRe (Chen et al., 2024) reconstructed foreground and background regions separately using 3DGS representations. However, these methods usually require extensive multi-view data acquisition and reconstruction. Recently, Bai et al. (2024) incorporated generative models for high-dynamic-range (HDR) environmental lighting estimation, making large-scale object insertion more practical. VISTA+ follows the foreground-background decomposition strategy but avoids large-scale multi-view reconstruction by combining single-view NVS for backgrounds, CAD-based foreground rendering, HDR-based lighting estimation, and image harmonization.

2.3. Synthetic data generation platforms

A wide range of simulation platforms has been developed for synthetic data generation in autonomous driving, including traditional graphics engines and specialized simulators such as CARLA (Dosovitskiy et al., 2017), AirSim (Shah et al., 2018), 51Sim-One, PreScan, and Virtual Test Drive (VTD). These platforms enable controllable scenario generation but often require labor-intensive manual scene construction. Neural-rendering

Just Accepted

simulators, such as MARS (Wu et al., 2023), OASim (Yan et al., 2024a), UniSim (Yang et al., 2023b), Street Gaussians (Yan et al., 2024b), and OmniRe (Chen et al., 2024), improve realism through automatic labeling, 3D reconstruction, generalization, and scene manipulation. Nevertheless, they typically depend on dense data collection, accurate pose estimation, and stable environmental lighting. These requirements also limit engineering scalability in real-world fleet deployments: pose-estimation errors can destabilize reconstruction, lengthy per-scene optimization must be repeated across routes, and illumination changes can reduce reconstruction robustness. Generative model-based simulators, including BEVControl (Yang et al., 2023a), DriveDreamer (Wang et al., 2023b), MagicDrive (Gao et al., 2023), and DriveArena (Yang et al., 2024), can synthesize realistic scenes conditioned on bird's-eye-view (BEV) maps, bounding boxes, and camera viewpoints, but they require large-scale training data and substantial computing resources. In contrast, VISTA (Amini et al., 2020, 2022) generates novel-view images from spatially nearby frames without expensive multi-view reconstruction. Motivated by this engineering advantage, VISTA+ adopts a single-view, feed-forward NVS strategy that avoids scene-specific multi-view reconstruction and neural-rendering optimization, while still using camera calibration and relative poses for reprojection, and further improves image realism through foreground inpainting, mesh-based background generation, CAD-based foreground rendering, and image harmonization. Table 1 summarizes the differences between VISTA+ and representative data-driven simulation frameworks.

Table 1 Similarities and differences between VISTA+ and representative data-driven simulation frameworks

Method	Foreground	Background NVS	Inference	Image composition
VISTA / VISTA 2.0	3D assets (VISTA 2.0)	Depth-based mesh	Online rendering	Rendered in scene
GeoSim	Neural reconstruction	Not specified	Offline pre-reconstruction	Image harmonization
Worldsheet	Not specified	Mesh	Online feed-forward	Not specified
VISTA+	CAD assets	Inpainted mesh	Online feed-forward	Image harmonization

3. Methodology

3.1. Synthetic data generation pipeline

To generate a scenario trajectory, road network files containing precise geographic coordinates are first imported into simulation software to produce trajectories for the ego vehicle and surrounding dynamic agents. The camera viewpoints of the ego vehicle and the poses of surrounding objects are then extracted from the generated trajectories and passed to the rendering layer for image generation.

For each target camera viewpoint, the pre-collected image with the minimum Euclidean distance to the target camera position is selected for NVS generation. A textured mesh is reconstructed from this image to synthesize the

Just Accepted

background view. Foreground images of dynamic objects are then generated using CAD assets rendered in Blender, where asset selection and rendering are guided by the object poses obtained from the trajectory. To improve photorealistic consistency, environmental lighting parameters are estimated from the generated background and applied to foreground rendering. Finally, the foreground and background layers are composited to generate synthetic images with corresponding labels.

In implementation, source images from the same camera sensor are prioritized for NVS generation (Fig. 1). Foreground and background regions are reconstructed independently and then fused. This strategy reduces photometric inconsistencies caused by differences in sensor properties, such as exposure duration, focal length, sensor type, and noise profile. Although the feasible NVS range is therefore constrained and multiple driving traces may be required for large viewpoint changes, this limitation is acceptable for many road scenarios where vehicle heading variations are relatively small. Moreover, because dynamic objects in the background are removed by inpainting, the proposed pipeline reduces artifacts caused by conflicts between inserted objects and the original dynamic objects. Geometry-aware object-insertion frameworks such as GeoSim (Chen et al., 2021) also explicitly consider object placement and occlusion.



Fig. 1 Single-view background NVS workflow.

Just accepted

3.2. Background novel view synthesis

The scene background is parameterized as a textured mesh, which enables efficient novel-view image synthesis from a single source image (Fig. 2). Specifically, the source frame spatially closest to the target camera viewpoint is retrieved from the pre-collected trajectory data. The background generation pipeline consists of three stages. First, an image inpainting model removes dynamic objects such as vehicles, pedestrians, and other non-smooth foreground entities according to semantic or bounding-box masks, producing a clean background image. Second, a mesh estimation network predicts grid deformation and depth, enabling 3D mesh reconstruction through differentiable rendering. Third, the final textured mesh is obtained by projecting the reconstructed mesh onto the inpainted background image and resampling the corresponding textures. Novel-view background images are then generated by rasterizing the textured mesh from the specified target camera viewpoint.

3.2.1. Foreground inpainting

In the preprocessing stage, a binary inpainting mask M is constructed from two-dimensional (2D) object bounding boxes in the image plane. Pixels within the masked regions are assigned a value of 1, while all other pixels are assigned 0. PowerPaint V2, the BrushNet-based version of PowerPaint (Zhuang et al., 2024; Ju et al., 2024), is then used to synthesize plausible background content in the masked regions. To guide the generative process, the positive prompt "empty scene, blurred background" is used, while negative prompts such as "vehicle, car, truck, bus, trailer, pedestrian, barrier, motorcycle, bicycle" are specified to suppress hallucinated foreground objects in the repaired regions.

Just Accepted

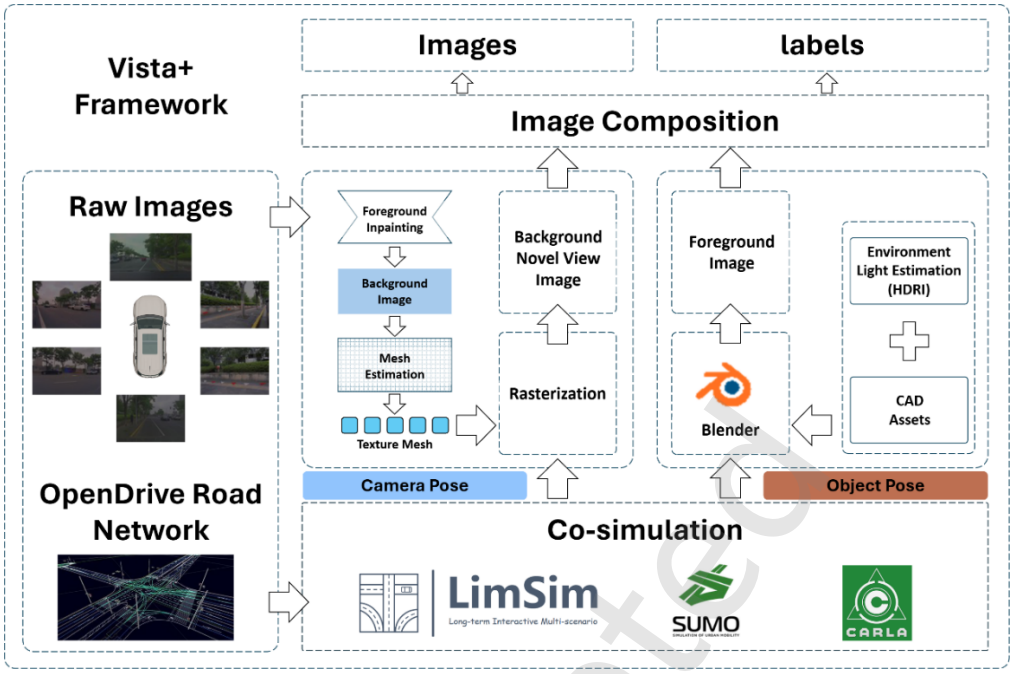


Fig. 2 Synthetic data generation pipeline of VISTA+.

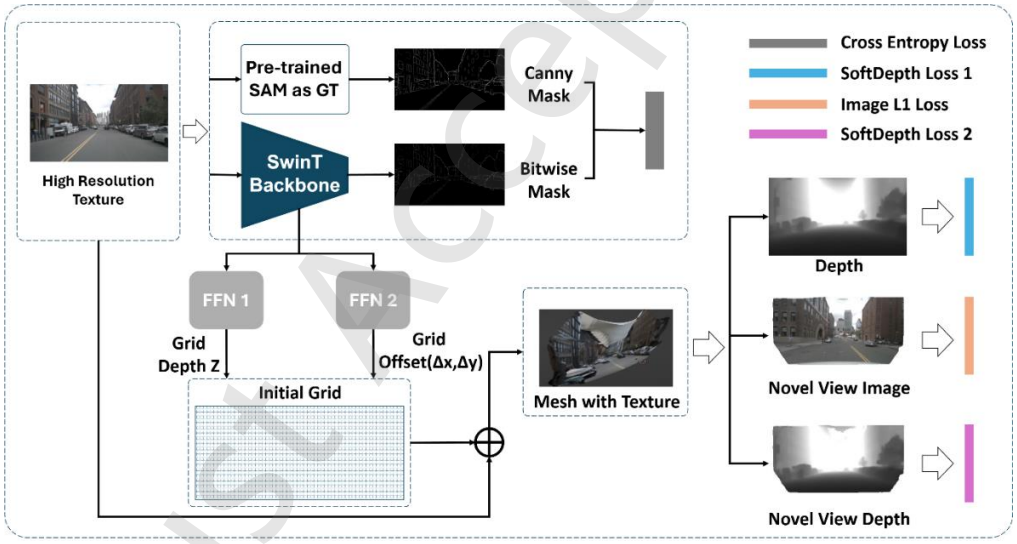


Fig. 3 Mesh estimation network and training losses. The SAM-derived supervision shown as "GT" provides pseudo-labels for the edge-prediction task.

Just Accepted

3.2.2. Mesh estimation

As shown in Fig. 3, the mesh prediction network estimates grid offsets (x, y) and depth values (z), which are combined with image textures for NVS. Specifically, the network predicts depth and image-plane offsets on a 57 × 100 mesh-vertex grid (height × width). The Swin Transformer (SwinT) (Liu et al., 2021) is adopted as the backbone because of its ability to capture local context and hierarchical visual features. Following the design principle of Worldsheet (Hu et al., 2021), the network processes an input image and regresses a depth map and mesh offsets through two feed-

forward network (FFN) heads. The predicted depth and offsets form an initial grid, which is combined with the input image to construct a preliminary textured mesh. The textured mesh can then be projected from arbitrary camera viewpoints to render corresponding depth maps and RGB images. The loss function contains three components: a source-view SoftDepth loss obtained by projecting the textured mesh back to the source viewpoint and comparing the rendered depth with the reference depth; a target-view SoftDepth loss computed after projecting the mesh to the reference target viewpoint; and an L1 photometric loss between the rendered RGB image and the corresponding real reference image. In this work, depth maps predicted by UniDepth (Piccinelli et al., 2024) are used as dense pseudo-labels for supervision. Light detection and ranging (LiDAR) data or other depth-sensing modalities can also be used when available. The total loss is:

$$\text{Total loss} = \lambda_1 \text{SoftDepth loss}_2 + \lambda_2 \text{SoftDepth loss}_1 + \lambda_3 \text{Image L1 loss} \quad (1)$$

In Eq. (1), SoftDepth loss 1 and SoftDepth loss 2 denote the source-view and target-view depth losses, respectively, consistent with Fig. 3. The coefficients λ1, λ2, and λ3 weight the target-view depth, source-view depth, and RGB losses, respectively. The loss is back-propagated through differentiable rendering implemented with PyTorch3D (Ravi et al., 2020).

The training process contains two stages. In the first stage, the SwinT backbone is pretrained using an edge prediction proxy task. Edge pseudo-labels are derived from the outputs of the pretrained Segment Anything Model (SAM) (Kirillov et al., 2023). This proxy task encourages the backbone to encode high-level geometric and structural priors. In the second stage, the pretrained SwinT weights are transferred to the depth and offset estimation task and further fine-tuned for mesh reconstruction. This two-stage strategy leverages structural knowledge learned from edge prediction to improve subsequent mesh estimation and NVS quality.

Novel-view synthesized images are supervised using a pixel-wise L1 loss between the predicted images and the corresponding ground-truth images. For depth supervision, a direct L2 loss is not ideal because vertices at different depths contribute unequally to the final rendered view. Therefore, a SoftDepth loss is introduced to adaptively account for depth-dependent rendering effects.

Just Accepted

$$\text{SoftDepth loss}=\begin{cases}\sqrt{fx^2+fy^2}\cdot\dfrac{|Z_{error}|}{Z(Z+Z_{error})} \\ (\text{if } Z<Z_{clamp} \text{ and } Z+Z_{error}<Z_{clamp}) \\ \\ 0 \quad \text{else}\end{cases} \tag{2}$$

Here, z_{error} denotes the difference between the reference and predicted depths, and z_{clamp} denotes the maximum distance that can be rendered.

3.2.3. Foreground image generation and image composition

Given the target six-degree-of-freedom (6-DoF) object pose and sensor intrinsics, an appropriate 3D model is retrieved from the CAD asset library and instantiated in Blender at the designated geospatial coordinates. The scene lighting is then configured, and the foreground rendering pipeline is executed.

Accurate environmental lighting is crucial for photorealistic foreground-background consistency. According to the lighting condition (e.g., sunny, rainy, or night) and scene type (e.g., highway, urban, or rural) of the generated background, an HDR image is selected from the HDR library as the base environment map. This base HDR image is then composed with sampled scene images to form a new HDR environment lighting source. The base HDR image is necessary because a single sampled image usually cannot provide sufficient sky illumination intensity or complete high-frequency lighting details. In practical applications, HDR images with relatively uniform illumination, such as cloudy-sky HDR maps, are preferred to ensure that foreground object details can be rendered reliably.

Object shadows are generated by HDR environmental lighting. Although the direction of sunlight can be estimated from Global Positioning System (GPS) and timestamp information, explicit sunlight is not introduced in this work. This design choice avoids physically inconsistent shadows in cases where foreground objects are located in building shadows and should not receive direct sunlight. Because determining when explicit sunlight should be introduced remains challenging, it is excluded from the current implementation.

The rendered foreground and reconstructed background layers are first fused through alpha compositing. To further reduce visual inconsistencies, an image harmonization step is applied to adjust the foreground appearance to the background. Image harmonization compensates for residual errors in environmental lighting estimation, allowing the pipeline to use efficient engineering-based lighting estimation while maintaining visual coherence.

Just Accepted

4. Experiments

The proposed approach is evaluated on NVS tasks and further assessed through synthetic-data augmentation for downstream 3D object detection.

4.1. Experimental setup

The mesh prediction network was trained on a single NVIDIA RTX A4000 graphics processing unit (GPU) with 16 GB memory and a batch size of 4 using the Adam optimizer (Kingma and Ba, 2015). Training followed the two-stage edge-prediction pretraining and mesh-reconstruction procedure described in Section 3.2.2. The network predicted a 57 × 100 mesh-vertex grid, and the total objective comprised the source-view SoftDepth loss, target-view SoftDepth loss, and RGB L1 loss defined in Section 3.2.2. Foreground rendering and image composition were conducted on a system equipped with an NVIDIA Quadro RTX 4000 GPU with 8 GB memory, a 2.1 GHz 16-core central processing unit (CPU), and 32 GB random-access memory (RAM).

4.2. Performance on the NVS task

Models were trained and evaluated on two widely used autonomous-driving datasets, KITTI (Geiger et al., 2012) and nuScenes (Caesar et al., 2020). For each dataset, one-tenth of the samples were held out for testing. For KITTI and the original nuScenes setting, frames were uniformly sampled from all available sequences, using frame t as the source view and frame t+1 as the target novel-view image. For the nuScenes (Inpainting) setting, annotated keyframes were selected, with keyframe t used as the input and keyframe t+1 used as the target view. Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM; Wang et al., 2004) were used as standard image-quality metrics.

We compare VISTA+ with three representative single-view NVS baselines: Flash3D (Szymanowicz et al., 2025), which uses a 3DGS representation; SynSin (Wiles et al., 2020), which uses a depth-based point-cloud representation; and the original Worldsheet (Hu et al., 2021), which uses a mesh representation. Quantitative results are summarized in Table 2, and qualitative comparisons are shown in Fig. 4. The results indicate that VISTA+ produces clearer and more realistic novel views, particularly for fine structures such as distant lane markings, which are more blurred in Flash3D and SynSin.

Just Accepted



Fig. 4 Comparison of NVS results produced by different models. The proposed method is labeled VISTA+.

The proposed method consistently outperforms representative 3DGS-, point-cloud-, and mesh-based single-view NVS baselines. This improvement suggests that the mesh representation is well suited to road scenes whose static backgrounds approximately satisfy smooth geometric continuity. The nuScenes (Inpainting) results show that VISTA+ remains effective when foreground removal and inpainting are included in the pipeline. However, because this setting uses a different sampling protocol from nuScenes (Original), the numerical difference between the two columns should not be attributed solely to inpainting. Inpainting is not evaluated on KITTI because the KITTI scenes used in this experiment contain relatively few foreground objects, and foreground removal provides limited additional benefit. The stronger performance on KITTI than on nuScenes further indicates that smoother backgrounds and fewer occlusions are more favorable for mesh-based NVS.

Table 2 NVS performance comparison on KITTI and nuScenes

Method	KITTI		nuScenes (Original)		nuScenes (Inpainting)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Flash3D	22.89	0.782	22.59	0.626	22.81	0.707
SynSin	20.91	0.713	20.52	0.593	21.14	0.631
Worldsheet	29.89	0.793	22.72	0.714	23.32	0.731
VISTA+ (SwinT; 57 × 100 mesh-vertex grid)	30.21	0.809	25.13	0.744	25.73	0.753

Note: nuScenes (Original) uses uniformly sampled consecutive frames, whereas nuScenes (Inpainting) uses consecutive annotated keyframes. Because the sampling protocols differ, the columns do not constitute a controlled inpainting ablation.

Only novel viewpoints are used for image-quality evaluation. This avoids overly optimistic results that would occur if the source viewpoint were included, because single-view mesh-based methods can trivially reproduce the

Just Accepted

original image at the source view. As shown in Table 2, VISTA+ achieves the best PSNR and SSIM on KITTI and under both nuScenes evaluation protocols. The numerical difference between the two nuScenes columns is not interpreted as the isolated effect of inpainting because the sampling protocols differ. These results support the effectiveness of the evaluated mesh-reconstruction pipeline, while also revealing that complex urban structures and frequent occlusions remain more challenging than smooth highway-like scenes.

4.3. Performance of synthetic data on downstream 3D object detection tasks

The original nuScenes dataset is used as the baseline training set. Throughout Tables 3 and 6, a p% synthetic-data setting is defined relative to the number of original training samples: $N_{syn} = (p/100)N_{orig}$, while all original samples are retained. Thus, the 50% setting contains N_{orig} original samples and $0.5N_{orig}$ synthetic samples, for a total of $1.5N_{orig}$ training samples. For Table 3, models are trained using the original data plus 50% additional synthetic data generated by VISTA+. Specifically, original objects are removed through inpainting according to ground-truth bounding boxes, replaced with objects of the same category, and rendered under random novel-view jitter. This evaluation measures the functional realism of the generated data: if the synthetic images improve perception performance, they provide useful visual and geometric variations rather than merely improving perceptual image quality.

Evaluation follows the nuScenes protocol (Caesar et al., 2020). The nuScenes detection score (NDS) and mean average precision (mAP) are accuracy metrics; mean average translation error (mATE), mean average scale error (mASE), mean average orientation error (mAOE), mean average velocity error (mAVE), and mean average attribute error (mAAE) are true-positive error metrics.

Table 3 3D detection results with synthetic data augmentation

Method	Training data	NDS	ΔNDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
Sparse4D	Original	0.541	–	0.436	0.633	0.279	0.363	0.317	0.177
Sparse4D	+50% synthetic	0.563	+0.022	0.459	0.574	0.256	0.334	0.324	0.177
StreamPETR	Original	0.540	–	0.432	0.581	0.272	0.413	0.295	0.195
StreamPETR	+50% synthetic	0.564	+0.024	0.454	0.516	0.248	0.373	0.303	0.194
BEVDet4D	Original	0.515	–	0.396	0.619	0.260	0.361	0.399	0.189
BEVDet4D	+50% synthetic	0.539	+0.024	0.422	0.563	0.238	0.324	0.402	0.188
BEVFormer	Original	0.517	–	0.416	0.673	0.274	0.372	0.394	0.198
BEVFormer	+50% synthetic	0.540	+0.023	0.433	0.600	0.244	0.338	0.385	0.198

Note: All original samples are retained in the +50% synthetic rows. ΔNDS is calculated relative to the original-data baseline for the same detector. Higher values are better for NDS and mAP, whereas lower values are better for mATE, mASE, mAOE, mAVE, and mAAE.

Just Accepted

The NDS is calculated using the standard nuScenes formulation:

$$\mathrm{NDS} = \frac{1}{10}\left[5\mathrm{mAP} + \sum_{\mathrm{mTP}\in\mathrm{TP}} (1 - \min(1, \mathrm{mTP}))\right] \tag{3}$$

Performance is assessed using several state-of-the-art 3D detection models, including Sparse4D (Lin et al., 2022), StreamPETR (Wang et al., 2023a), and BEV-based models such as BEVDet4D (Huang and Huang, 2022) and BEVFormer (Li et al., 2022).

As shown in Table 3, adding synthetic data improves the overall NDS and mAP of all four detectors. The NDS increases by 0.022 for Sparse4D, 0.024 for StreamPETR, 0.024 for BEVDet4D, and 0.023 for BEVFormer, while the corresponding mAP gains are 0.023, 0.022, 0.026, and 0.017, respectively. These gains are accompanied by consistent reductions in mATE (0.056–0.073), mASE (0.022–0.030), and mAOE (0.029–0.040), indicating improved object localization, scale estimation, and orientation estimation. By contrast, the changes in mAVE and mAAE are small, suggesting that the principal benefits of VISTA+ augmentation arise from improved spatial and visual diversity rather than temporal or attribute cues. The consistent improvements across sparse-query-based, temporal, and BEV-based detectors further indicate that the generated samples are not specific to a single model architecture.

For the six multi-camera examples shown in Fig. 5, an image-based estimate indicates that border-connected missing regions occupy approximately 4.4% of the nominal image area on average, ranging from 0.6% to 8.0% across views. This is an illustrative, figure-specific estimate rather than a dataset-wide statistic, because the missing fraction depends on camera geometry, scene depth, and viewpoint displacement. Although the aggregate results in Table 3 show that the generated data provide an overall benefit, they do not establish that missing peripheral regions are harmless. Such regions may reduce contextual information and valid supervision for objects projected near image boundaries and may disproportionately affect small, distant objects when their image footprints overlap an invalid region. Future work will quantify the missing-area distribution from renderer visibility masks and evaluate detection performance by object range and border proximity.

4.4. Result analysis

To better understand the performance and limitations of VISTA+, qualitative results and failure cases are analyzed below.

Just Accepted

4.4.1. Image generation

Fig. 5 presents several generated novel-view images, showing that VISTA+ can synthesize realistic perspectives with consistent geometry. Fig. 6 further illustrates examples generated using the nearest textured mesh. The six background images are attached to the same ego vehicle and therefore follow the vehicle motion. The black vehicle is rendered multiple times according to the camera pose of each image to generate corresponding foreground images.

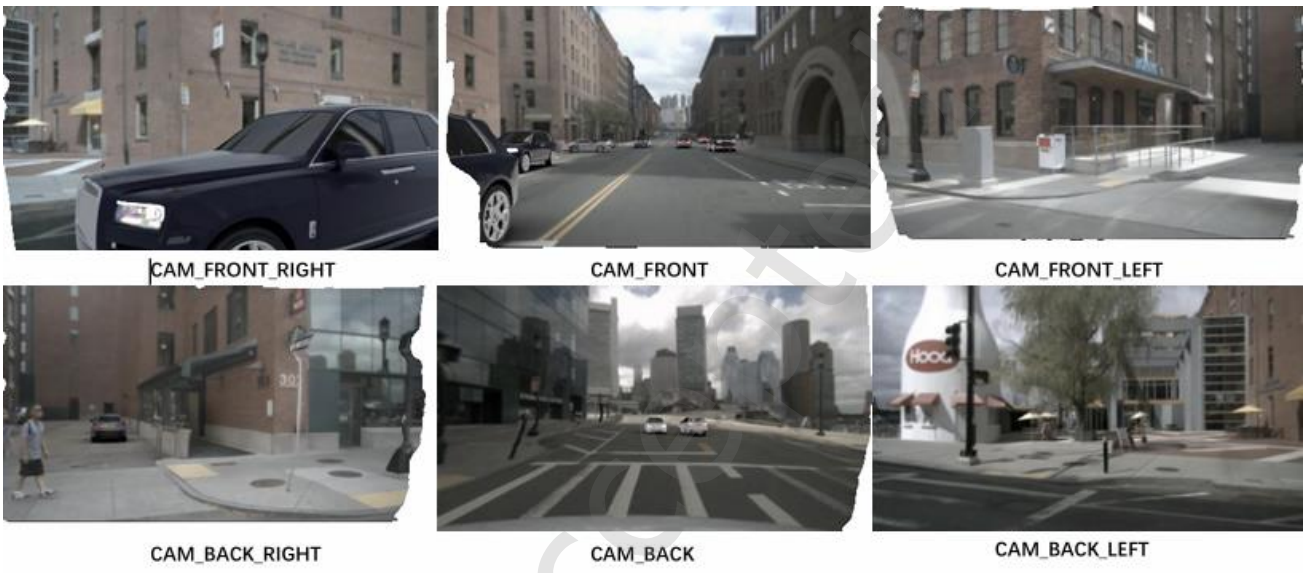


Fig. 5 Novel-view multi-camera images with vehicles rendered by Blender.

4.4.2. Video generation

VISTA+ is also evaluated for video synthesis. Fig. 7 shows a sequence in which the front-view camera is shifted laterally by 1 m and 2 m in both directions. The generated video exhibits smooth transitions and plausible geometry, demonstrating that VISTA+ can handle moderate viewpoint changes. However, distortions appear around elongated roadside objects, especially poles and barriers, where the reconstructed structures become stretched. Fig. 8 illustrates this failure case alongside an example generated using ideal depth supervision. Because these examples depict different scenes, the comparison does not isolate depth estimation as the cause of the distortion.

4.4.3. Non-ideal NVS reconstruction

In addition to geometric distortion, mesh reconstruction quality degrades under adverse lighting conditions, such as nighttime, rain, and fog. In these cases, generated images may lose color fidelity, texture details, and sharpness, weakening the final rendering quality. In complex urban scenes where large foreground objects such as vehicles or pedestrians heavily occlude the background, the inpainting module may also fail to recover realistic background details. These cases reveal the limitations of inpainting-based background recovery when occlusion is severe.

Just Accepted

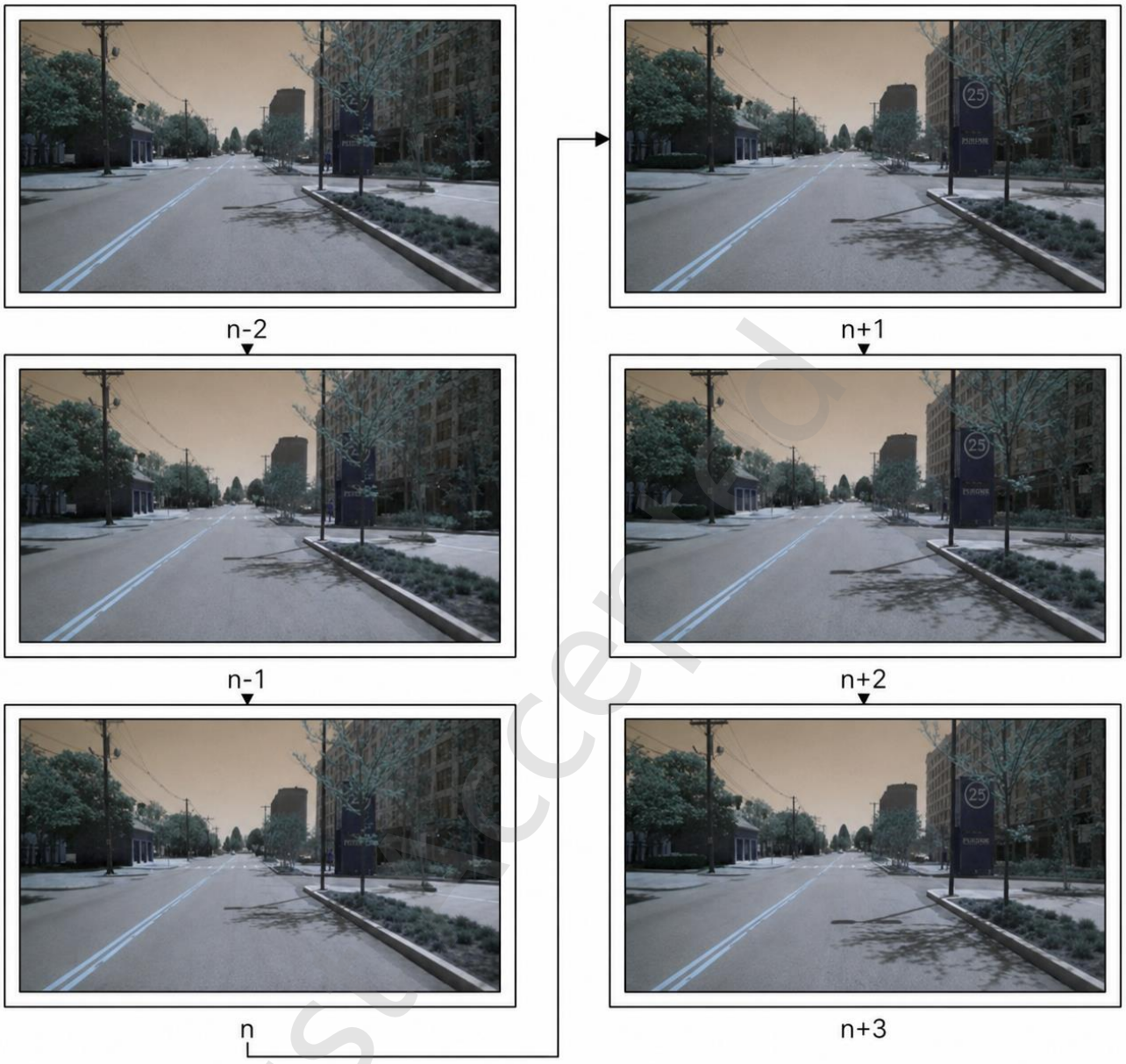


Fig. 6 Images generated with the nearest textured mesh. Frames n-2 and n-1 are generated from frame n, while frames n+1 and n+2 are generated from frame n+3. The uniform white spacing around the panels is schematic layout spacing and does not represent the fraction of missing image content.



Fig. 7 Video generation by the proposed method. The front-view camera is shifted horizontally by ±1 m and ±2 m.

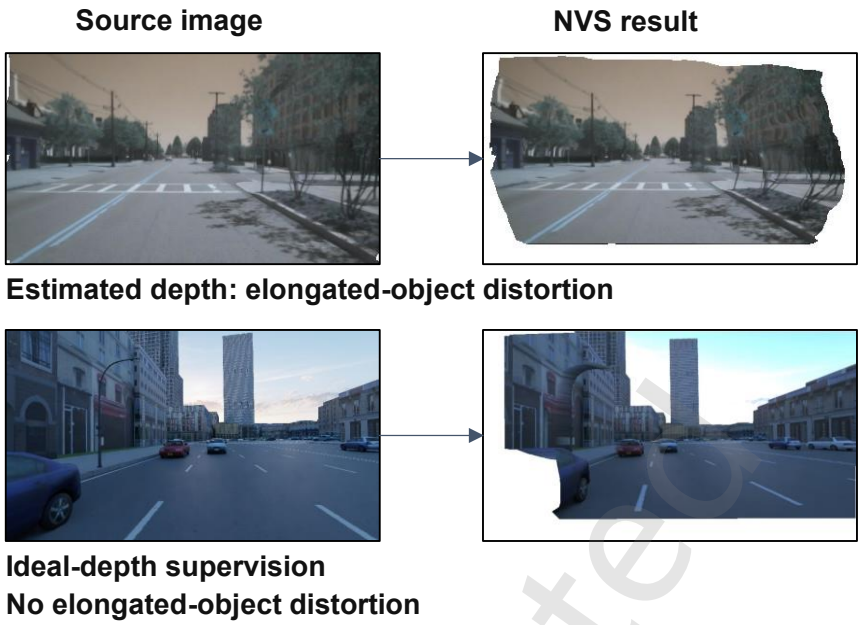


Fig. 8 Distortion around elongated objects. The estimated-depth example (top) shows stretched thin structures, while the ideal-depth example (bottom) illustrates a different scene. This is a qualitative illustration, not a controlled comparison of depth supervision.

4.5. Ablation study

4.5.1. Different depth estimation methods

Several depth estimation approaches are evaluated on KITTI to analyze their influence on NVS quality while minimizing the impact of foreground objects. The compared depth sources include (1) sparse depth points derived from binocular stereo, (2) sparse depth obtained from LiDAR measurements, and (3) dense depth predictions generated by UniDepth (Piccinelli et al., 2024). As shown in Table 4, dense depth predicted by a state-of-the-art pretrained model yields higher reconstruction quality than sparse stereo or LiDAR depth. This result suggests that dense monocular depth estimation is sufficient for the proposed framework in practical applications.

Table 4 NVS performance comparison using different depth estimation methods

Method	KITTI	
	PSNR↑	SSIM↑
Sparse depth from binocular stereo	28.34	0.721
Sparse depth from LiDAR	28.42	0.734
Dense depth from pretrained UniDepth	30.13	0.803

The lower performance of the sparse-depth settings is likely attributable to their limited image-plane coverage, which

Just Accepted

leaves many mesh vertices weakly constrained or dependent on interpolated supervision, particularly near depth discontinuities. Calibration and projection errors can further misalign sparse measurements with image edges and introduce geometric distortions during rendering. This result therefore reflects the sparse depth configurations evaluated here rather than an inherent disadvantage of LiDAR; accurately calibrated, high-density LiDAR could provide more complete supervision and narrow the observed gap.

4.5.2. Impact of model modifications

To further evaluate the training strategy, we conduct ablation experiments on architectural and optimization choices, including the backbone network, the two-stage training paradigm, and joint prediction of depth and vertex offsets.

As summarized in Table 5, SwinT outperforms the ResNet50 and ResNet101 convolutional neural network (CNN) backbones (He et al., 2016) for vertex offset prediction, indicating the advantage of attention-based hierarchical feature modeling. Pretraining the backbone on an edge prediction proxy task further improves performance by providing structural priors for mesh estimation. Jointly training the network to predict depth and vertex offsets also contributes to more accurate and robust mesh reconstruction.

Table 5 NVS performance with different network architectures and training settings

Method	nuScenes	
	PSNR↑	SSIM↑
ResNet50	24.82	0.718
ResNet101	25.03	0.733
SwinT	25.07	0.736
SwinT + edge-prediction pretraining	25.10	0.742
SwinT + edge-prediction pretraining + joint depth-offset training	25.13	0.744

4.5.3. Impact of synthetic data proportion on 3D detection performance

Table 6 reports the effect of different synthetic-data proportions on downstream detection performance. The Sparse4D model was trained on 8 NVIDIA A100 GPUs with a batch size of 8 using the AdamW optimizer (Loshchilov and Hutter, 2019). The initial learning rates for the backbone and other parameters were set to 2×10^{-5} and 2×10^{-4}, respectively, with cosine annealing decay. Training required approximately 3 days, and the inference speed was 15 frames per second. The highest NDS and mAP are achieved when 50% synthetic data is added to the original training set, while larger proportions provide smaller gains; at 200%, mAP falls below the original-data baseline in Table 3, suggesting that excessive synthetic data may introduce a domain gap.

Just Accepted

Table 6 Evaluation results on nuScenes with different proportions of synthetic training data

Method	Data	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
Sparse4D	+30% synthetic	0.550	0.448	0.620	0.271	0.352	0.324	0.177
	+50% synthetic	0.563	0.459	0.574	0.256	0.334	0.324	0.177
	+100% synthetic	0.552	0.448	0.618	0.271	0.346	0.308	0.177
	+200% synthetic	0.543	0.424	0.602	0.267	0.328	0.319	0.179

4.6. Limitations and future work

Although the proposed framework achieves promising results, several limitations require further investigation, including adverse illumination, severe foreground occlusion, incomplete field-of-view boundaries, and real-time deployment constraints.

First, in adverse lighting conditions such as nighttime, rain, or fog, the loss of high-frequency details reduces the fidelity of mesh reconstruction. To mitigate this limitation, the pipeline incorporates a generative image translation model, img2img-turbo (Parmar et al., 2024), built upon Stable Diffusion Turbo (Sauer et al., 2024), to transform well-lit scenes into realistic adverse-weather or low-light images.

Second, when large foreground objects heavily occlude the background, current inpainting algorithms may fail to recover the underlying geometry and texture, resulting in artifacts. A possible solution is multi-path reconstruction, in which additional images with clearer background visibility are collected and used to improve reconstruction quality.

Third, incomplete image margins remain unavoidable when the target viewpoint moves beyond the field of view of the source image. Future work will explore generative completion models to reconstruct missing margins. Alternatively, 360-degree panoramic cameras could naturally reduce this boundary limitation.

VISTA+ can also be combined with trajectories generated by simulation software such as SUMO (Alvarez Lopez et al., 2018), LimSim (Wen et al., 2023), and CARLA (Dosovitskiy et al., 2017) for closed-loop end-to-end simulation. However, public autonomous-driving datasets are often designed for object detection and contain dense dynamic objects, making it difficult for inpainting models to recover clean and temporally consistent backgrounds. Future work will therefore collect or identify datasets more suitable for route-based reconstruction, improve inpainting with generative world models, and optimize the inference speed of VISTA+ for real-time applications. In addition, current validation mainly focuses on regular road segments. In complex intersections or irregular road structures, large viewpoint changes may introduce geometric distortions. To address this issue, future versions will incorporate a confidence mechanism based on depth consistency or reprojection error to automatically filter low-

Just Accepted

quality synthetic views. Safety measures derived from vehicle spacing at urban roundabouts (Bhavsar et al., 2025) could complement image-quality and perception metrics in future closed-loop evaluations.

5. Conclusions

This study presents VISTA+, a practical synthetic driving-scene generation framework that couples mesh-based single-view NVS with CAD-based foreground rendering and foreground-background composition. On KITTI, VISTA+ achieved a PSNR of 30.21 dB and an SSIM of 0.809, exceeding the strongest baseline by 0.32 dB and 0.016, respectively. In the original nuScenes setting, VISTA+ achieved a PSNR of 25.13 dB and an SSIM of 0.744, compared with 22.72 dB and 0.714 for the strongest baseline; in the separately sampled nuScenes (Inpainting) setting, it reached 25.73 dB and 0.753. Adding synthetic data equal to 50% of the original training set improved NDS by 0.022–0.024 and mAP by 0.017–0.026 across four 3D object detectors. These results demonstrate that VISTA+ improves both image-synthesis quality and downstream detection without requiring large-scale multi-view scene reconstruction. The current method remains limited by missing peripheral regions and geometric distortion under large viewpoint changes, but it provides an engineering-efficient solution for repeated-route applications such as autonomous-driving development fleets, fixed-route buses, and robotaxis.

Just Accepted

Author contributions

Song Zhang and Ruibo Wang: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **Jian Li, Wenxiu Shi, Huiping Zhu, Ruo Jia, and Ziyu Lin:** Writing – review & editing, Methodology. **Haiquan Liu:** Supervision, Funding Acquisition.

Acknowledgements

This work was supported by the "U35 Pei Miao Funding Program" of Changping District, Beijing (No. CHPU35202504003), the National Natural Science Foundation of China (No. 52572334), and the Independent Research Project of the State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University (No. ZZ-GG-20250403).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Replication and data availability

The experiments in this study use the publicly available KITTI and nuScenes datasets. The replication data associated with this study are available on ETS-Data under Dataset ID **ETS-Data-26-00097**.

Declaration of competing interest

The authors declare that they have no competing interests relevant to the content of this article.

References

Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., Rother, C., 2018. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision 126, 961-972.

Alvarez Lopez, P., Wiessner, E., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y. P., et al., 2018. Microscopic traffic simulation using SUMO. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2575-2582. https://doi.org/10.1109/ITSC.2018.8569938.

Amini, A., Gilitschenski, I., Phillips, J., Moseyko, J., Banerjee, R., Karaman, S., et al., 2020. Learning robust control policies for end-to-end autonomous driving from data-driven simulation. IEEE Robotics and Automation Letters 5, 1143-1150.

Amini, A., Wang, T. H., Gilitschenski, I., Schwarting, W., Liu, Z., Han, S., et al., 2022. VISTA 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles. In: 2022 International Conference on Robotics and Automation (ICRA), 2419-2426.

Bai, C., Shao, Z., Zhang, G., Liang, D., Yang, J., Zhang, Z., et al., 2024. Anything in any scene: Photorealistic video object insertion. https://arxiv.org/abs/2401.17509.

Bhavsar, Y. M., Zaveri, M. S., Raval, M. S., Zaveri, S. B., 2025. Evaluating defensive driving behaviour based on safe distance between vehicles: A case study using computer vision on UAV videos at urban roundabout. Multimodal Transportation 4(3), 100227. https://doi.org/10.1016/j.multra.2025.100227.

Blender Foundation, 2024. Blender. https://www.blender.org.

Just Accepted

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., et al., 2020. nuScenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11621-11631.

Chen, Y., Rong, F., Duggal, S., Wang, S., Yan, X., Manivasagam, S., et al., 2021. GeoSim: Realistic video simulation via geometry-aware composition for self-driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7230-7240.

Chen, Z., Yang, J., Huang, J., de Lutio, R., Martinez Esturo, J., Ivanovic, B., et al., 2024. OmniRe: Omni urban scene reconstruction. https://arxiv.org/abs/2408.16760.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An open urban driving simulator. In: Conference on Robot Learning, 1-16.

Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D. Y., et al., 2023. MagicDrive: Street view generation with diverse 3D geometry control. https://arxiv.org/abs/2310.02601.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 3354-3361. https://doi.org/10.1109/CVPR.2012.6248074.

Gómez, J. L., Silva, M., Seoane, A., Borrás, A., Noriega, M., Ros, G., et al., 2025. All for one, and one for all: UrbanSyn Dataset, the third Musketeer of synthetic driving scenes. Neurocomputing 637, 130038. https://doi.org/10.1016/j.neucom.2025.130038.

Hammami, A., Borsos, A., Sándor, Á. P., 2025. How realistic a bicycle simulator can be? - A validation study. Multimodal Transportation 4(1), 100193. https://doi.org/10.1016/j.multra.2025.100193.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

He, Y., Lin, H., Yang, L., Liu, Y., 2025. Generative models for the evolution of transportation systems. Journal of Traffic and Transportation Engineering (English Edition) 12(6), 1784-1803. https://doi.org/10.1016/j.jtte.2024.03.010.

Hu, R., Ravi, N., Berg, A. C., Pathak, D., 2021. Worldsheet: Wrapping the world in a 3D sheet for view synthesis from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 12528-12537.

Huang, J., Huang, G., 2022. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. https://arxiv.org/abs/2203.17054.

Just Accepted

Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., et al., 2023. VAD: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8340-8350.

Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q., 2024. BrushNet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. arXiv preprint arXiv:2403.06976.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics 42(4), Article 139, 1-14. https://doi.org/10.1145/3592433.

Kingma, D. P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations. https://arxiv.org/abs/1412.6980.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015-4026.

Klinghoffer, T., Philion, J., Chen, W., Litany, O., Gojcic, Z., Joo, J., et al., 2023. Towards viewpoint robustness in bird's eye view segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8515-8524.

Kurse, T. K., Gebresenbet, G., Daba, G. F., Tefera, N. T., 2025. Experimental determination of factors causing crashes involving automated vehicles. Multimodal Transportation 4(1), 100186. https://doi.org/10.1016/j.multra.2024.100186.

Li, J., Feng, Z., She, Q., Ding, H., Wang, C., Lee, G. H., 2021. MINE: Towards continuous depth MPI with NeRF for novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 12578-12588.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., et al., 2022. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: European Conference on Computer Vision, 1-18. https://doi.org/10.1007/978-3-031-20077-9_1.

Lin, H., Chen, H., Wang, R., Zhang, S., Liu, Y., 2026a. A NeRF-based road-surface geometry fusion method for camera-only 3D object detection. IEEE Transactions on Intelligent Vehicles, 1-10. https://doi.org/10.1109/TIV.2026.3733707.

Lin, H., Liu, Y., Li, S., Qu, X., 2023. How generative adversarial networks promote the development of intelligent transportation systems: A survey. IEEE/CAA Journal of Automatica Sinica 10(9), 1781-1796. https://doi.org/10.1109/JAS.2023.123744.

Just Accepted

Lin, H., Shi, W., Huang, H., Zhuang, D., Zhang, S., Liu, Y., Qu, X., Zhao, J., 2026b. Risk-controllable multi-view diffusion for driving scenario generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 5169-5178.

Lin, X., Lin, T., Pei, Z., Huang, L., Su, Z., 2022. Sparse4D: Multi-view 3D object detection with sparse spatial-temporal fusion. https://arxiv.org/abs/2211.10581.

Liu, J. Y., Chen, Y., Yang, Z., Wang, J., Manivasagam, S., Urtasun, R., 2023. Real-time neural rasterization for large scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8416-8427.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al., 2021. Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations. https://openreview.net/forum?id=Bkg6RiCqY7.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2022. NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65(1), 99-106. https://doi.org/10.1145/3503250.

Niu, L., Cong, W., Liu, L., Hong, Y., Zhang, B., Liang, J., et al., 2021. Making images real again: A comprehensive survey on deep image composition. arXiv:2106.14490. https://arxiv.org/abs/2106.14490.

Ost, J., Mannan, F., Thuerey, N., Knodt, J., Heide, F., 2021. Neural scene graphs for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2856-2865.

Parmar, G., Park, T., Narasimhan, S., Zhu, J. Y., 2024. One-step image translation with text-to-image models. https://arxiv.org/abs/2403.12036.

Piccinelli, L., Yang, Y. H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F., 2024. UniDepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10106-10116.

Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W. Y., Johnson, J., et al., 2020. Accelerating 3D deep learning with PyTorch3D. https://arxiv.org/abs/2007.08501.

Sauer, A., Lorenz, D., Blattmann, A., Rombach, R., 2024. Adversarial diffusion distillation. In: European Conference on Computer Vision, 87-103.

Shah, S., Dey, D., Lovett, C., Kapoor, A., 2018. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: Hutter, M., Siegwart, R. (Eds.), Field and Service Robotics. Springer Proceedings in Advanced Robotics 5, 621-635. https://doi.org/10.1007/978-3-319-67361-5_40.

Just Accepted

Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S., 2025. SparseDrive: End-to-end autonomous driving via sparse scene representation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), 8795-8801.

Szymanowicz, S., Insafutdinov, E., Zheng, C., Campbell, D., Henriques, J. F., Rupprecht, C., Vedaldi, A., 2025. Flash3D: Feed-forward generalisable 3D scene reconstruction from a single image. In: 2025 International Conference on 3D Vision (3DV), 670-681.

Tsirikoglou, A., Kronander, J., Wrenninge, M., Unger, J., 2017. Procedural modeling and physically based rendering for synthetic data generation in automotive applications. https://arxiv.org/abs/1710.06270.

Tucker, R., Snavely, N., 2020. Single-view view synthesis with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 551-560.

Tulsiani, S., Tucker, R., Snavely, N., 2018. Layer-structured 3D scene inference via view synthesis. In: European Conference on Computer Vision, 302-317.

Tulsiani, S., Zhou, T., Efros, A. A., Malik, J., 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2626-2634.

Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X., 2023a. Exploring object-centric temporal modeling for efficient multi-view 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3621-3631.

Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J., 2023b. DriveDreamer: Towards real-world-driven world models for autonomous driving. arXiv:2309.09777. https://arxiv.org/abs/2309.09777.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600-612. https://doi.org/10.1109/TIP.2003.819861.

Wen, L., Fu, D., Mao, S., Cai, P., Dou, M., Li, Y., Qiao, Y., 2023. LimSim: A long-term interactive multi-scenario traffic simulator. In: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), 1255-1262.

Wiles, O., Gkioxari, G., Szeliski, R., Johnson, J., 2020. SynSin: End-to-end view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7467-7477.

Wimbauer, F., Yang, N., Rupprecht, C., Cremers, D., 2023. Behind the scenes: Density fields for single view reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9076-9086.

Just Accepted

Wrenninge, M., Unger, J., 2018. Synscapes: A photorealistic synthetic dataset for street scene parsing. https://arxiv.org/abs/1810.08705.

Wu, Z., Liu, T., Luo, L., Zhong, Z., Chen, J., Xiao, H., et al., 2023. MARS: An instance-aware, modular and realistic simulator for autonomous driving. arXiv:2307.15058. https://arxiv.org/abs/2307.15058.

Yan, G., Pi, J., Guo, J., Luo, Z., Dou, M., Deng, N., et al., 2024a. OASim: An open and adaptive simulator based on neural rendering for autonomous driving. arXiv:2402.03830. https://arxiv.org/abs/2402.03830.

Yan, Y., Lin, H., Zhou, C., Wang, W., Sun, H., Zhan, K., et al., 2024b. Street Gaussians: Modeling dynamic urban scenes with Gaussian splatting. arXiv:2401.01339. https://arxiv.org/abs/2401.01339.

Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K., 2023a. BEVControl: Accurately controlling street-view elements with multi-perspective consistency via BEV sketch layout. https://arxiv.org/abs/2308.01661.

Yang, X., Wen, L., Ma, Y., Mei, J., Li, X., Wei, T., et al., 2024. DriveArena: A closed-loop generative simulation platform for autonomous driving. arXiv:2408.00415. https://arxiv.org/abs/2408.00415.

Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W. C., Yang, A. J., Urtasun, R., 2023b. UniSim: A neural closed-loop sensor simulator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1389-1399.

Zhang, S., Lin, H., Wang, M., Wei, B., Liu, Y., Qu, X., 2026. A dynamic prompting and scenario generation method for autonomous driving perception via large-model optimization. Transportation Research Part C: Emerging Technologies 188, 105672. https://doi.org/10.1016/j.trc.2026.105672.

Zhu, Y., Wang, S., Zhong, W., Shen, N., Li, Y., Wang, S., et al., 2026. A survey on large language model-powered autonomous driving. Engineering 63, 190-203. https://doi.org/10.1016/j.eng.2025.07.038.

Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K., 2024. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In: European Conference on Computer Vision, 195-211.

Author biography



Song Zhang received his M.S. degree at the department of Precision Instruments and Mechanical in Tsinghua University in 2007. Currently, he works at Z-one Technology Co., Ltd., SAIC Motor Corporation. His research interests focus on deep learning perception algorithms for autonomous driving applications, generative models, data synthesis, reinforcement learning, and 3D reconstruction algorithms.

Just Accepted



Ruibo Wang earned his M.S. degree in science (computer science) at the University of Melbourne, Melbourne, Australia, in 2021. Currently, he is an engineer at Z-one Technology Co., LTD. His research interests include 3D reconstruction, Semantic synthesis and LLM research.



Jian Li received the M.S. degree from the School of Information Science and Engineering in Southest University. He is now working in the Z-one Technology Co., Ltd., SAIC Motor Corporation as an algorithm engineer. His research interests include autonomous driving perception algorithms and data synthesis algorithms.



Wenxiu Shi received her Master's degree from Shenzhen University in 2018. She currently serves as a Multimodal Large Model Algorithm Engineer at Z-one Technology Co., Ltd., SAIC Motor Corporation. Her research concentrates on generative AI, 3D visual reconstruction, and multimodal perception models, with a focus on data synthesis, algorithm innovation, and practical applications of AIGC technology in autonomous driving scenarios.



Huiping Zhu received the M.S. degree from School of Electronic Information and Electrical Engineering of Shanghai Jiao Tong University in 2019. He is currently an Algorithm Engineer in Z-one Technology Co., Ltd., SAIC Motor Corporation. His main research interests are deep learning perception algorithms for autonomous driving and generative models.

Just Accepted