

Data-driven Camera and Lidar Simulation Models for Autonomous Driving: A Review from Generative Models to Volume Renderers

Hamed Haghighi¹, Xiaomeng Wang¹, Hao Jing¹, and Mehrdad Dianati²

Abstract—Perception sensors, particularly camera and Lidar, are key elements of Autonomous Driving Systems (ADS) that enable them to comprehend their surroundings for informed driving and control decisions. Therefore, developing realistic simulation models for these sensors is essential for conducting effective simulation-based testing of ADS. Moreover, the rise of deep learning-based perception models has increased the utility of sensor simulation models for synthesising diverse training datasets. The traditional sensor simulation models rely on computationally expensive physics-based algorithms, specifically in complex systems such as ADS. Hence, the current potential resides in data-driven approaches, fuelled by the exceptional performance of deep generative models in capturing high-dimensional data distribution and volume renderers in accurately representing scenes. This paper reviews the current state-of-the-art data-driven camera and Lidar simulation models and their evaluation methods. It explores a spectrum of models from the novel perspective of generative models and volume renderers. Generative models are discussed in terms of their input-output types, while volume renderers are categorised based on their input encoding. Finally, the paper illustrates commonly used evaluation techniques for assessing sensor simulation models and highlights the existing research gaps in the area.

Index Terms—data-driven, sensor simulation, deep generative models, GANs, diffusion models, volume rendering, neural radiance fields, 3D Gaussian splatting, image synthesis, 3D point cloud synthesis, camera, Lidar, and autonomous driving systems.

I. INTRODUCTION

Safety is crucial in Autonomous Driving Systems (ADS), given the potentially severe consequences of system failures, as evidenced by recent Uber and Tesla crashes [1], [2], which highlight the need for stringent testing protocols. Physical testing of ADS, though valuable, is challenging due to the time, labour, and cost involved, and requires extensive miles to statistically validate safety [3]. Moreover, certain dangerous scenarios may not be feasible or ethical for real-world testing. Conversely, virtual testing in simulation environments offers several advantages, including the efficient simulation of extensive miles in a short period, testing safety-critical scenarios without physical damage, and modelling complex and costly-to-recreate traffic scenarios. In the domain of ADS applications, perception sensors, including cameras and Lidar, play a pivotal role. These sensors monitor the

surrounding environment, detecting moving objects such as vehicles, cyclists, pedestrians, and stationary objects such as traffic lights and road signs. It is critical to test ADS with the realistic performance of the perception sensors, especially in challenging environments where their performances are likely to be degraded. Hence, development of realistic sensor simulation models becomes vital, facilitating extensive testing and validation in simulated environments and contributing to the overall safety and reliability of ADS.

The surge in deep learning-based models for ADS, especially in perception applications [4], has led to a substantial demand for annotated sensory datasets to train these models. For this reason, numerous sensory datasets have been recorded from real driving scenes [5] and have been annotated by human labour in recent years. However, the process of collecting and annotating real-world datasets is costly and presents challenges such as privacy concerns and safety hazards. As a solution, many researchers have turned to simulation environments to generate synthetic datasets. Simulation frameworks enable the rapid creation of extensive sensory data with ground-truth annotations and the generation of edge-case scenarios without posing safety hazards. These synthetic datasets are either used in conjunction with real datasets to train perception models or employed independently with domain adaptation techniques. In both scenarios, realistic simulation of perception sensors plays a significant role in enhancing the performance of downstream perception tasks.

The literature on sensor simulation models presents two fundamental approaches: physics-based and data-driven techniques [6]. Physics-based methods involve explicit simulations of sensor-related physical phenomena, relying on intricate hand-crafted formulations for approximations. For instance, Liu et al. [7] introduced a high-fidelity physics-based camera model for autonomous driving, incorporating components that precisely simulate light propagation, surface materials, camera lens, and aperture. Although capable of generating high-fidelity sensor images, such systems require extensive computations. In contrast, data-driven models have gained popularity in recent years to address the complexity of physics-based models. Unlike physics-based approaches, data-driven models leverage statistical models to implicitly uncover underlying relations by learning from data. The growing availability of real-world recorded perception sensory datasets, coupled with the success of deep generative models in synthesising high-dimensional sensory data and accuracy of volume renderers in 3D scene representation, has led to a rapid expansion of

¹H. Haghighi, X. Wang, and H. Jing are with WMG, University of Warwick, Coventry, U.K. (Corresponding author: Hamed.Haghighi@warwick.ac.uk)

²M. Dianati is with the School of Electronics, Electrical Engineering and Computer Science at Queen’s University of Belfast and the WMG at the University of Warwick.

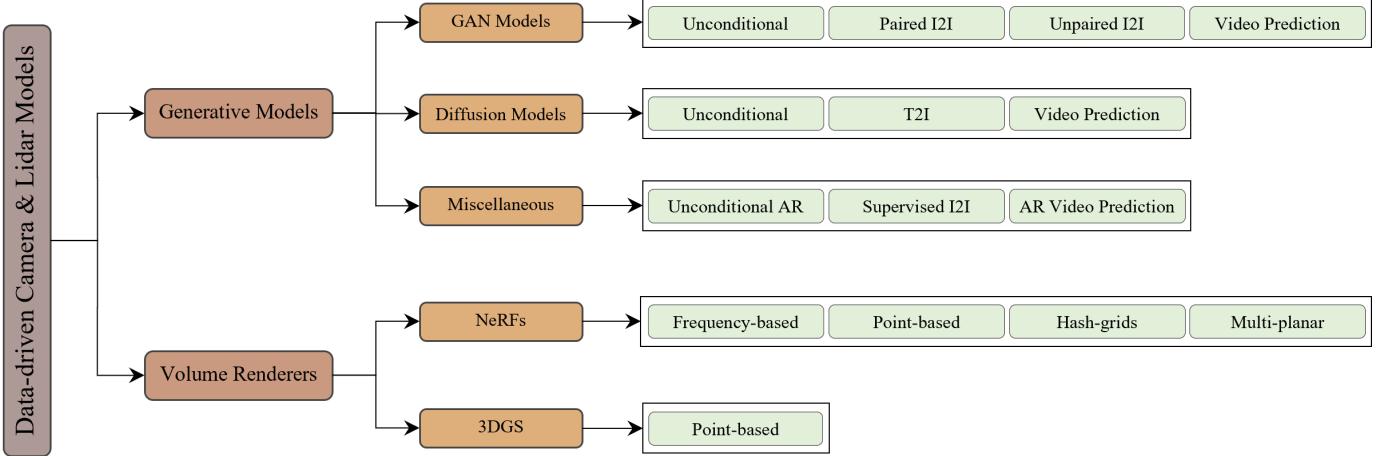


Fig. 1: Categorisation of data-driven camera and Lidar simulation models for ADS.

the literature on data-driven models.

In this article, we review the State-Of-The-Art (SOTA) data-driven camera and Lidar simulation models, and their evaluation techniques. We propose a novel perspective exploring methods from the standpoint of generative models and volume renderers. Generative models focus on modelling sensory data through implicit distribution approximators such as Generative Adversarial Networks (GANs) [8], denoising diffusion Models [9], or Auto-Regressive (AR) models. On the other hand, volume renderers adopt a more explicit approach, utilising learning-based models, such as Neural Radiance Fields (NeRFs) [10] or 3D Gaussian Splatting (3DGS) [11] models, for scene representation and ray-marching. We further categorise the generative models based on their input-output data into unconditional, Image-to-Image (I2I) translation, Text-to-Image (T2I) translation and video prediction models. Additionally, volume renderers are further categorised based on their input encoding methods into frequency-based, hash-grids, point-based and multi-planar approaches. The categorisation of data-driven camera and Lidar simulation models is shown in Fig. 1. Furthermore, we review the existing evaluation methods for data-driven sensor simulation models, exploring both qualitative and quantitative approaches.

In the context of ADS simulation, several literature review papers have investigated various aspects, including driving simulation frameworks, synthetic datasets, and sensor simulation approaches. For instance, Rosique et al. [12] conducted a review of perception systems and simulators for ADS, emphasising the characteristics of sensor hardware and simulators used in vehicle tests, game engines, and robotics. Kang et al. [13] provided an overview of public driving datasets and virtual testing environments, focusing on accessible virtual testing environments for closed-loop ADS testing. Schlager et al. [6] conducted a study reviewing perception sensor models, categorising radar, Lidar, and camera models based on fidelity levels. In a recent work [14], the authors concentrated on reviewing digital camera components and their simulation approaches in the context of ADS and robotics. Despite the coverage of various sensor simulation methods, there is a noticeable gap in the discussion of SOTA data-driven methods,

particularly those based on generative models and volume renderers. Furthermore, existing research has not adequately investigated sensor simulation evaluation approaches, a crucial aspect for virtual verification and validation of ADS. The summary of our paper’s contributions is as follows:

- A comprehensive literature review of data-driven camera and Lidar simulation models is carried out, with a specific emphasis on the latest techniques.
- A novel perspective on data-driven sensor simulation models is discussed, exploring both implicit approaches such as generative models and more explicit models such as volume renderers.
- A detailed explanation and categorisation of evaluation approaches for sensor simulation models are provided.

The structure of this paper is as follows: Section II provides background information on data-driven simulation models, emphasising both generative models and volume renderers. Sections III, IV, and V delve into various generative models, including GAN-based models, diffusion models and miscellaneous, respectively. Sections VI and VII address volume renderers, focusing on NeRFs and 3DGS models, respectively. Section VIII outlines methods for the evaluation of simulation models. Finally, Section IX provides concluding remarks, identifies research gaps, and suggests directions for future research.

II. BACKGROUND

Data-driven approaches model sensory data by leveraging large real-world datasets to train models that predict or simulate new sensory outputs with high accuracy. Unlike traditional methods that rely on manual formulas and exact physics-based simulations, data-driven methods abstract patterns directly from data, enabling them to dynamically adapt to complex and variable real-world conditions. This capability significantly enhances the scalability and realism of simulations.

In the landscape of data-driven modelling, generative models and volume rendering models represent two distinct paradigms. Generative models, including Generative Adversarial Networks (GANs) [8] and denoising diffusion models [9],

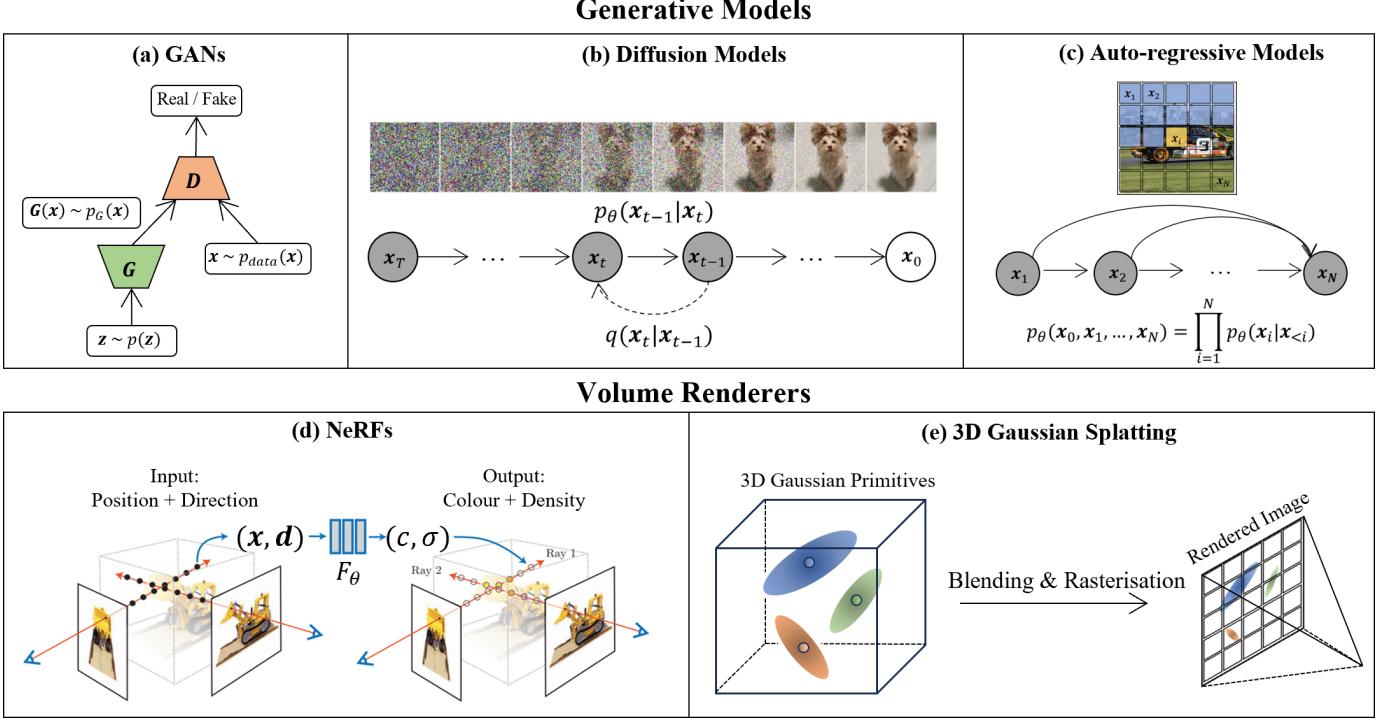


Fig. 2: An overview of data-driven models, including generative models and volume renderers, that are widely used for camera and Lidar simulation in ADS. These data-driven models contain generative approaches such as (a) GANs [8], (b) denoising diffusion models [9], and (c) auto-regressive models, while volume renderers include (d) NeRFs [10] and (e) 3D Gaussian splatting models [11]. The car image in (c) is sourced from ImageNet [15].

learn statistical data distributions to implicitly generate new, plausible data instances. They excel in producing diverse scenarios without explicitly defining every possible condition, making them ideal for generating diverse sensory data. In contrast, volume renderers such as Neural Radiance Fields (NeRFs) [10] provide a representation of the environment. They construct highly detailed and accurate 3D reconstructions by mapping spatial coordinates to visual properties, which can be rendered from any viewpoint. This distinction paves the way for a detailed exploration of each method's specific techniques and advantages, focusing on how they address different needs in simulation.

A. Generative Models

Generative models attempt to learn the distribution of real-world data and generate new instances that mimic this distribution without the need for explicit scene geometry or physical properties. In the context of sensor simulation, generative models aim to approximate a target probability distribution $p_{\text{data}}(\mathbf{x})$ derived from a set of observed sensory data $\mathbf{x} \in \mathbb{R}^N$, where N is the dimensionality of the sensor output. The goal is to learn a parameterised model distribution $p_\theta(\mathbf{x})$ that closely represents $p_{\text{data}}(\mathbf{x})$. Training a generative model involves adjusting the parameters θ of the model such that $p_\theta(\mathbf{x})$ becomes an effective approximation of $p_{\text{data}}(\mathbf{x})$. Generative models may employ a variety of strategies depending on their architecture and the specific training regime. Techniques such as adversarial training in GANs or the iterative refinement

in diffusion models are used to align $p_\theta(\mathbf{x})$ with $p_{\text{data}}(\mathbf{x})$, emphasising model flexibility in capturing and reproducing the complex statistical properties of sensory data.

1) Generative Adversarial Networks (GANs): Generative Adversarial Networks (GANs) [8] is a form of generative model, which consists of a generator (G) and a discriminator (D). The generator G attempts to generate data that mimics real-world data, learning to map from a latent space $z \in \mathbb{R}^M (M < N)$ to the data space, aiming to match the real data distribution $p_{\text{data}}(\mathbf{x})$. In contrast, the discriminator D evaluates the realism of samples, tasked with distinguishing real data from that generated by G , as shown in Fig. 2 (a).

The interaction between G and D is formulated as a min-max game, represented by the following equation:

$$\min_{G} \max_{D} \mathcal{L}(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{z \sim p(z)} \log(1 - D(G(z))) \quad (1)$$

Advantages of GANs include their ability to generate high-quality, realistic images, which makes them particularly useful for applications requiring high visual fidelity, such as automotive simulation. However, challenges with GANs include their training instability. The adversarial nature of their training can lead to issues such as non-convergence and mode collapse, where G fails to produce diverse outputs and instead generates repetitive or overly similar samples. Moreover, the sensitivity of GANs to the settings of hyper-parameters requires careful tuning to achieve optimal performance and stability.

2) *Diffusion Models*: Denoising diffusion probabilistic models [9], referred to as Diffusion Models (DMs), represent another class of generative models that operate by progressively adding and then removing noise to generate data. These models simulate the forward process where noise is incrementally added to the data \mathbf{x}_0 from the real data distribution $p_{data}(\mathbf{x}_0)$, resulting in a gradually noisier dataset over a sequence of steps $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. The reverse process involves a model that learns to denoise this data, effectively reconstructing the data back towards its original form (See Fig. 2 (b)). The mathematical formulation of DMs is centred around the concept of reversing the noise addition process. The generative process models the conditional probability of recovering a previous state x_{t-1} from a noisier state x_t , described by:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \quad (2)$$

where μ_θ and Σ_θ are functions parameterised by the model that estimate the mean and variance needed to denoise x_t .

The optimisation of DMs is typically approached by minimising a variational lower bound, often resulting in a simplified objective such as the mean squared error between the denoised and original data across the diffusion steps. The loss function generally used can be expressed as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [| | \epsilon - \epsilon_\theta(\mathbf{x}_t, t) | |^2], \quad (3)$$

where ϵ is the noise added at each step, and ϵ_θ is the noise model predicted by the network.

Diffusion models are known for their ability to generate high-quality and diverse outputs. Their stability in training is a significant advantage over other generative models, as they are less prone to issues like mode collapse, where the generator fails to capture the diversity of the dataset. However, a notable challenge with diffusion models is their computational efficiency. The iterative nature of the reverse process, which requires multiple network evaluations to generate a single sample, makes them computationally intensive compared to models that generate outputs in a single forward pass.

3) *Auto-regressive Models*: Auto-regressive (AR) models form another important category of generative models used for data-driven sensor simulation. These models generate data dimensions sequentially, where each data dimension is conditioned on the previous ones. In an AR model, the probability of a data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ is factorised into a product of conditional probabilities:

$$p(\mathbf{x}) = \prod_{i=1}^N p(\mathbf{x}_i | \mathbf{x}_{<i}), \quad (4)$$

where $\mathbf{x}_{<i}$ represents all the previous dimensions before i^{th} dimension (See Fig. 2 (c)).

Training an AR model involves maximising the likelihood of the training data under the model. This is typically done by minimising the negative log-likelihood:

$$\mathcal{L}(\theta) = - \sum_{i=1}^T \log p_\theta(\mathbf{x}_t | \mathbf{x}_{<t}). \quad (5)$$

Auto-regressive models can be implemented using architectures such as Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs) [16], and transformers [17]. These architectures effectively capture long-range dependencies and complex patterns in the data. One key advantage of AR models is their ability to generate data with high fidelity and temporal coherence, making them well-suited for tasks such as video prediction and time-series forecasting. However, the sequential nature of AR models can lead to slower inference times, as each data point must be generated step-by-step.

B. Volume Renderers

Volume rendering is a significant technique in computer graphics for visualising volumetric data, which involves simulating the propagation of light through a three-dimensional space. Contrary to generative models that indirectly encode the complexity of an environment by learning its data distribution, volume rendering models use explicit spatial information. They reconstruct scenes by integrating the interaction of light with the volumetric attributes of the environment, such as density and colour at every point in a three-dimensional space. This explicit approach is especially advantageous for simulations requiring precise optical and physical realism, as it can accurately depict how light interacts within complex environments.

The foundational principle of volume rendering can be captured by the volume rendering equation, which computes the contribution of light absorbed and emitted as it travels through a volume. The general formulation is:

$$C(\mathbf{r}) = \int_{t_0}^{t_1} T(t) \sigma(\mathbf{x}(t)) \mathbf{c}(\mathbf{x}(t), \mathbf{d}) dt, \quad (6)$$

where $C(\mathbf{r}) \in \mathbb{R}$ denotes the colour accumulated along ray $\mathbf{r} \in \mathbb{R}^3$, $\sigma(\mathbf{x}(t)) \in \mathbb{R}$ is the volume density at point $\mathbf{x}(t) \in \mathbb{R}^3$, $\mathbf{c}(\mathbf{x}(t), \mathbf{d}) \in \mathbb{R}^3$ represents the emitted colour at that point dependent on direction $\mathbf{d} \in \mathbb{R}^3$, and $T(t) \in \mathbb{R}$ is the transmittance, representing the light's attenuation from the start of the ray at $t_0 \in \mathbb{R}$ to $t \in \mathbb{R}$, calculated as $\exp(-\int_{t_0}^t \sigma(\mathbf{x}(s)) ds)$. This equation allows for detailed simulation of lighting effects through different materials, making volume rendering particularly effective.

1) *Neural Radiance Fields (NeRFs)*: Neural Radiance Fields (NeRFs) [10] are a type of volume renderers, that utilise continuous scene representations; They optimise a Multi-Layer Perceptron (MLP) to simulate light interactions within 3D environments through volumetric ray-marching. The fundamental NeRF formulation integrates radiance along a camera ray, calculating colour as a function of accumulated light and material properties at each point along the ray:

$$C(\mathbf{r}) = \int_{t_0}^{t_1} T(t) \sigma_\theta(\mathbf{x}(t)) \mathbf{c}_\theta(\mathbf{x}(t), \mathbf{d}) dt. \quad (7)$$

This integration considers both the emitted color \mathbf{c}_θ and the density σ_θ , which are output by a MLP (See Fig. 2 (d)). The transmittance $T(t)$ describes how much light survives without being scattered or absorbed, crucial for the realistic rendering

TABLE I: Comparison of widely used data-driven simulation methods in ADS.

Method	Category	3D Scene Representation	Output Realism	Inference Speed	Long-range Modelling	Training Stability	Output Diversity
Generative Models	GANs [8]	×	***	***	***	***	***
	Diffusion Models [9]	×	***	***	***	***	***
	AR Models	×	**	***	***	***	***
Volume Renderers	NeRFs [10]	✓	***	**	**	**	**
	3DGS [11]	✓	**	***	***	***	***

of materials like fog, smoke, or translucent objects. NeRFs excel at producing photo-realistic images and are capable of synthesizing new views from limited datasets. However, they require extensive computational resources for training and inference and are best suited for static scenes.

2) *3D Gaussian Splatting*: 3D Gaussian Splatting (3DGS) [11] introduces a novel approach to volume rendering that merges the explicit representation advantages of traditional methods, such as meshes and points, with the continuous scene modelling of NeRFs. By employing a tile-based splatting mechanism with anisotropic 3D Gaussian, this method achieves the flexibility required for real-time rendering (see Fig. 2 (e)). The 3DGS utilise 3D points characterised by specific features such as colour $\mathbf{c}_i \in \mathbb{R}$, weight $\sigma_i \in \mathbb{R}$, mean $\mathbf{p}_i \in \mathbb{R}^3$, and covariance $\Sigma_i \in \mathbb{R}^{3 \times 3}$ to represent the scene. It then projects the points on the image plane and transforms the features based on this projection to calculate the final image intensity. The Gaussian weight in the image plane for a point $\mathbf{x} \in \mathbb{R}^2$ is as:

$$\sigma_i(\mathbf{x}) = \frac{1}{|\mathbf{J}_i^{-1}|} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^T (\mathbf{J}_i \Sigma_i \mathbf{J}_i^T)^{-1} (\mathbf{x} - \mathbf{x}_i)\right), \quad (8)$$

where $\mathbf{J}_i \in \mathbb{R}^{2 \times 3}$ is the Jacobian matrix that transforms the 3D point \mathbf{p}_i during its projection onto the 2D image plane, resulting in \mathbf{x}_i . This transformation adapts the Gaussian shape to accommodate the camera's perspective effects.

The final image intensity at a pixel position \mathbf{x} in the 2D plane is derived by integrating the contributions from all points weighted by their respective Gaussian distributions:

$$c(\mathbf{x}) = \frac{\sum_{i=1}^N \sigma_i(\mathbf{x}) \cdot \mathbf{c}_i}{\sum_{i=1}^N \sigma_i(\mathbf{x})}. \quad (9)$$

This formulation ensures that the spatial and depth cues from the 3D scene are accurately represented in the 2D image, providing a realistic rendering of the scene according to the camera optics and geometry.

C. Comparative Analysis of Methods

This section comprehensively compares the discussed data-driven simulation methods, highlighting their key characteristics and capabilities. Table I summarises these attributes using a scoring system ranging from 1 to 3, represented by '*' symbols. A higher score indicates the superior performance of the model in the respective aspect.

1) *3D Scene Representation*: Volume renderers, specifically NeRFs and 3DGS, inherently represent complex 3D scenes, synthesising novel views and capturing spatial relationships. In contrast, generative models such as GANs, diffusion models, and auto-regressive models typically focus on capturing data distributions rather than accurately representing 3D scenes. While powerful and versatile, these generative models lack the inherent 3D representation capabilities that make volume renderers particularly suited for novel view synthesis tasks.

2) *Output Realism and Inference Speed*: Diffusion models and NeRFs demonstrate superior performance in terms of output realism, producing highly detailed and accurate results. However, this quality often comes at the cost of computational efficiency. 3DGS offers an attractive compromise, providing slightly less realism but with significantly faster inference speed than NeRFs. GANs, while potentially sacrificing some realism compared to diffusion models, excel in inference speed. This makes them one of the fastest methods for generating outputs, particularly useful in applications where real-time performance is crucial.

3) *Long-range Modelling*: Auto-regressive models excel in capturing and modelling long-range dependencies in data, allowing them to generate coherent sequences and maintain consistency over extended outputs. Diffusion models and volume renderers demonstrate moderate capabilities in this aspect, balancing local detail with broader context. GANs, however, typically struggle with long-range consistency, potentially limiting their effectiveness in tasks requiring extended coherence.

4) *Training Stability*: Diffusion models offer the highest level of training stability among the compared methods, demonstrating resilience to hyperparameter and architecture changes. NeRFs and 3DGS also demonstrate good stability during the training process, contributing to their effectiveness in 3D scene representation. GANs, while powerful, are known for their training instabilities and potential issues like mode collapse. This can make the training process more challenging and potentially less predictable. Among generative models, auto-regressive models generally offer moderate training stability.

5) *Output Diversity*: Diffusion models and auto-regressive models excel in generating diverse outputs, capable of producing a wide range of variations in their results. GANs can also generate diverse outputs, although they may sometimes suffer from mode collapse, potentially reducing diversity. NeRFs and 3DGS, being primarily focused on scene representation, typically offer lower output diversity. This is because they are usually trained on and reproduce specific scenes, prioritising

TABLE II: Summary of unconditional GAN-based Models.

Model	Sensor	Year	Description	Datasets
LidarGAN [18]	Lidar	2018	Synthesises high-quality Lidar scans using deep generative models	KITTI [19]
SB-GAN [20]	Camera	2019	Generates semantic label maps first, then synthesises the final image	Cityscapes [21]
Volokitin et al. [22]	Camera	2020	Separates image generation into layout prediction and detailed image synthesis	Cityscapes [21]
Semantic Palette [23]	Camera	2021	Uses class proportions to guide the generative process for scene elements	Cityscapes [21]
DUSTy [24]	Lidar	2021	Uses a noise-aware GAN framework to handle dropped points in Lidar scans	KITTI [19], MPO [24]
Dusty-2 [25]	Lidar	2022	Focuses on data-level domain transfer for Lidar range images	Raw KITTI [19]
Urban-StyleGAN [26]	Camera	2023	Enables detailed manipulation of urban scene images with high fidelity	Cityscapes [21]

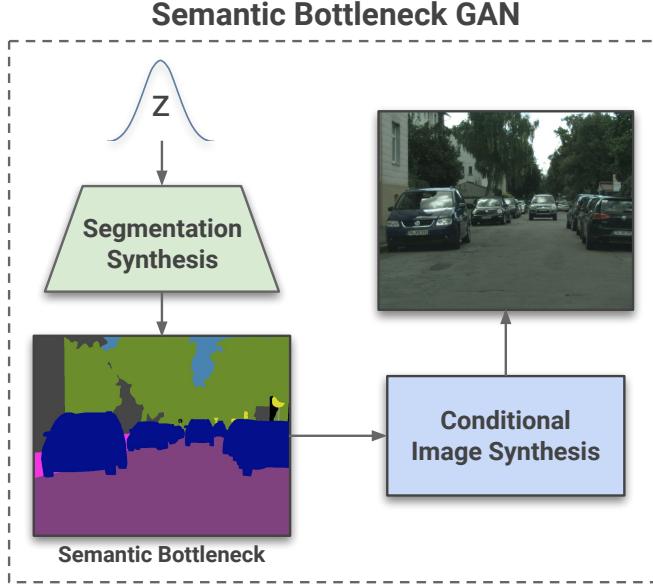


Fig. 3: Two-stage synthesis pipeline of the Semantic Bottleneck-GAN [20], an unconditional GAN model for RGB image synthesis.

accuracy over variation.

The comparison of these different characteristics highlights the distinct strengths and weaknesses of each method. Diffusion models stand out for their balance of high-quality outputs, diversity, and training stability. Auto-regressive models demonstrate particular strength in long-range modelling and output diversity. GANs offer the advantage of rapid inference, albeit with potential instabilities. NeRFs and 3DGS excel in 3D scene representation and realism, though they may have limitations in their broader generative capabilities. When selecting among these methods, the decision should be guided by the specific requirements and priorities of the task at hand, carefully weighing factors such as output quality, processing speed, diversity of results, and the dimensionality of the required representations.

III. GENERATIVE MODELS – GANS

GAN-based models constitute the largest category of data-driven approaches for sensor simulation in ADS. The GAN framework relies on the dynamic interaction between two key components: a generator that synthesises data and a discriminator that assesses its realism. Initially designed for

unconditional data generation from random noise, GANs have evolved to encompass conditional variants. These variants incorporate additional inputs to guide the data generation process, enabling different applications from both paired and unpaired image-to-image translation to spatio-temporal video prediction. The subsequent sections will explore each category of GAN models for ADS in greater detail.

A. Unconditional Models

Unconditional GAN models aim to approximate the distribution of observed sensory data without any external guidance or conditions. These models primarily concentrate on capturing the diversity of the data distribution while generating realistic samples. While not typically used directly for sensor simulation in ADS, they serve as foundational frameworks for more complex models. The summary of unconditional GAN models is provided in Table II. Below, we separately dig deep into unconditional models for camera RGB images and Lidar point cloud synthesis.

1) *RGB Image synthesis*: Synthesising RGB images from scratch is challenging due to the multi-object and highly diverse nature of driving scenes. To address this, unconditional models often separate the process into generating the semantic layout of the scene first, followed by the RGB image. The following section chronologically reviews papers in this domain, all of which conduct their experiments on the Cityscapes [21] dataset.

The Semantic Bottleneck GAN (SB-GAN) [20] and the decomposed synthesis Model [22] introduce novel approaches for generating urban scenes through a two-step process. They employ a two-step method where a semantic label map is first generated unconditionally and then used to guide the synthesis of the final image (as shown in Fig. 3), enhancing the generation of complex, high-resolution images with coherent global structures.

Progressing further, the introduction of the Semantic Palette Model [23] marks a significant innovation by using class proportions to guide the generative process. This approach allows users to specify the desired distribution of scene elements, providing unprecedented control over the class composition of generated images. This model greatly enhances the flexibility and practicality of GANs, making them more adaptable to varied application requirements.

The most recent advancement, Urban-StyleGAN [26] enables the generation and detailed manipulation of urban scene

TABLE III: Summary of paired I2I translation GAN-based models (all used for camera RGB image synthesis).

Model	Year	Description	Datasets
Pix2Pix [27]	2016	Aligns input semantic label maps with corresponding images using cGAN and L1 losses	Cityscapes [21]
Pix2PixHD [28]	2017	Incorporates multi-scale generator and discriminator for high-resolution image synthesis	Cityscapes [21]
SPADE [29]	2019	Introduces spatially-adaptive normalisation layers for semantic map-based modulation	Cityscapes [21]
CC-FPSE [30]	2019	Predicts layout-to-image conditional convolution kernels for image generation	Cityscapes [21]
SEAN [31]	2019	Allows per-region style control using Semantic Region-Adaptive Normalization	Cityscapes [21]
SurfelGAN [32]	2020	Combines texture-mapped surfels and GANs for realistic camera image generation	Waymo [33]
OASIS [34]	2020	Redesigns the discriminator to use semantic label maps directly as ground-truth	Cityscapes [21]
ECGAN [35]	2020	Uses edge information for enhancing semantic consistency and detail in image synthesis	Cityscapes [21]
Robusta [36]	2023	Enhances robustness of semantic segmentation models with high-quality perturbed images	Cityscapes [21]

images with remarkable fidelity and control. This model incorporates various techniques for managing and manipulating the latent space, providing tools for precise image editing that push the boundaries of realism in generated urban landscapes.

2) *Lidar Point Cloud Synthesis*: The LidarGAN model [18] adapts deep generative models such as VAEs [37] and GANs to synthesise high-quality Lidar scans by unrolling them into 2D point maps. This method not only generates realistic samples but also learns meaningful latent representations of the data. By augmenting the 2D signal with absolute positional information, the model enhances robustness against noisy and incomplete input data. Building on this foundation, the DUSTy model [24] introduces a noise-aware GAN framework that tackles the challenge of dropped points in Lidar scans caused by measurement uncertainty. DUSTy incorporates a differentiable sampling framework to simulate dropout noises, thereby enhancing the generation of high-quality Lidar data from incomplete observations.

Further advancing, the Dusty-2 model [25] proposes a generative approach for Lidar range images, focusing on data-level domain transfer and addressing issues such as inconsistent angular resolution and missing properties. This model uses an implicit image representation-based GAN coupled with a differentiable ray-drop effect to enhance the fidelity and diversity of the generated data. Its effectiveness is showcased in tasks such as upsampling, data restoration, and sim-to-real semantic segmentation.

B. Paired Image-to-Image-Translation Models

Paired Image-to-Image translation (I2I) models are a specific type of conditional model where the input image (condition) and the desired output (ground-truth) are directly paired in the training dataset. In the context of ADS, these models are typically trained using pairs, such as a semantic segmentation layout with its corresponding RGB image, known as semantic image synthesis. While it is feasible to train paired I2I models with supervised learning, incorporating the GAN framework significantly enhances the realism of the synthesised images. For example, the Pix2Pix model [27] aligns each input semantic label map with its corresponding image using both the

conditional GAN loss $\mathcal{L}_{\text{cGAN}}$ and the L1 loss \mathcal{L}_{L1} , defined as:

$$\mathcal{L}_{\text{Pix2Pix}}(G, D) = \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{\text{L1}}(G) \quad (10)$$

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\log D(\mathbf{x}, \mathbf{y})] + \mathbb{E}_{\mathbf{x}}[\log(1 - D(\mathbf{x}, G(\mathbf{x})))]$$

$$\mathcal{L}_{\text{L1}}(G) = \mathbb{E}_{\mathbf{x}, \mathbf{y}}[\|\mathbf{y} - G(\mathbf{x})\|_1],$$

where $\mathbf{x} \in X$ represents the input image, $\mathbf{y} \in Y$ represents the ground-truth image, and λ is a weighting factor that balances two losses. The adversarial loss drives the generator to produce images indistinguishable from real ones, while the L1 loss reduces blurring, promoting sharpness. The following reviews focus on I2I models for camera RGB image synthesis, all of which have been evaluated on the Cityscapes dataset unless otherwise specified. The summary of paired I2I translation models based on GANs is provided in Table III.

Building on Pix2pix model, Pix2PixHD [28] model addresses the limitations of the original Pix2Pix by incorporating a multi-scale generator and discriminator architecture, as shown in Fig. 4. This allows for the synthesis of high-resolution images with more detail and texture from semantic maps, thus overcoming the constraints of generating high-quality results from low-resolution inputs.

The innovation continues with the SPADE (Spatially-Adaptive Normalisation) [29] model, which introduces spatially-adaptive normalisation layers that modulate activations based on the input's semantic segmentation map. This effectively preserves and utilises semantic information throughout the synthesis process, setting a new standard for photo-realistic and high-resolution image synthesis. Advancing further, the CC-FPSE [30] model introduces the concept of predicting layout-to-image conditional convolution kernels, allowing the semantic layout to directly influence the image generation process. By learning spatially-varying convolution kernels based on the input semantic layout, this model offers more effective control over the details and alignment of the synthesised images with their corresponding semantic layouts.

The SEAN [31] model takes this a step further by introducing Semantic Region-Adaptive Normalisation (SEAN), which allows for per-region style control. Building on SPADE's foundation, SEAN enables the use of different style images for each semantic region, resulting in more detailed and realistic image synthesis. The integration of these styles into the normalisation layers allows SEAN to produce high-quality images with fine control over regional styles.

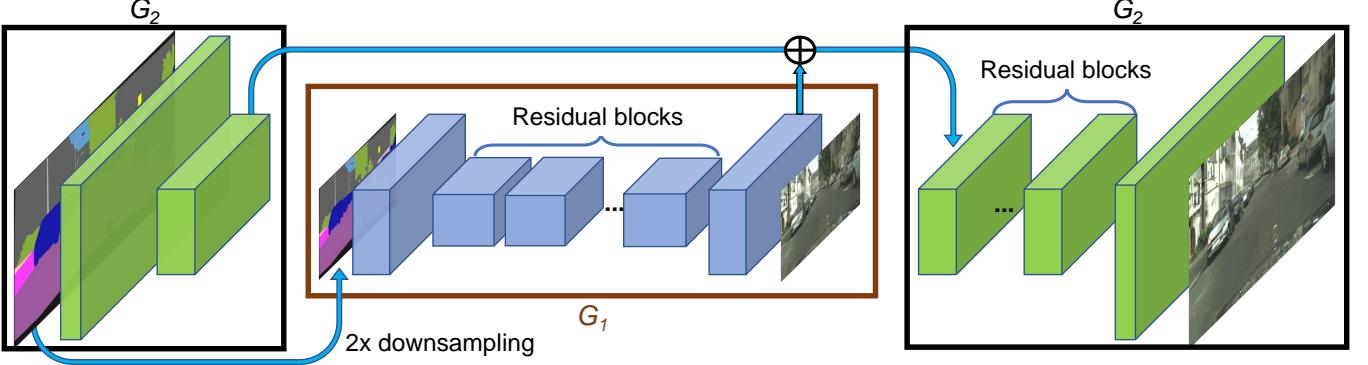


Fig. 4: The multi-scale synthesis process of pix2pixHD [28], a paired I2I model based on GANs.

ECGAN [35] tackles the challenges of synthesising local details and structures in semantic image synthesis by using edge information as an intermediate representation. This model introduces an attention-guided edge transfer module and a novel contrastive learning method to enhance semantic consistency and detail. By focusing on both local and global semantic information, ECGAN significantly improves the quality of synthesised images, particularly in capturing fine details and small objects that previous models often miss.

SurfelGAN [32] brings a novel approach to generating realistic sensor data for autonomous driving simulations. By combining texture-mapped surfels, a 3D reconstruction method, with GANs, SurfelGAN creates realistic camera images from novel viewpoints, evaluated on the Waymo Open dataset [33]. This method vastly improves the realism of synthetic data compared to traditional simulation methods using gaming engines, enhancing the training and testing of ADS.

The OASIS [34] model simplifies the GAN framework for semantic image synthesis by redesigning the discriminator to use semantic label maps directly as ground truth for training. This innovation eliminates the need for a VGG-based perceptual loss, which is typically used to enhance the quality of synthesised images, thereby simplifying the training process and improving the fidelity of the output images.

Finally, the Robusta [36] model focuses on improving the robustness of semantic segmentation models by generating high-quality, perturbed images using a novel conditional GAN architecture. Robusta utilises a two-stage GAN architecture: a coarse generator for handling label-to-image translation and a fine generator for improving image quality. The inclusion of attention layers and spatially-adaptive normalisation enables better handling of anomalies and distribution shifts, significantly enhancing the robustness of segmentation models in real-world scenarios.

C. Unpaired Image-to-Image Translation Models

Unpaired I2I translation models are a distinct type of conditional GANs where the input and output in the training dataset do not have a direct correspondence. These models typically require a consistency loss between the input and output domains, along with a GAN loss during training. For example, the pioneering CycleGAN [38] introduced the cycle-consistency loss \mathcal{L}_{cyc} , which can be defined as follows:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} [\|\mathbf{x}\|_1 + \mathbb{E}_{\mathbf{y} \sim p_{data}(\mathbf{y})} [\|G(F(\mathbf{y})) - \mathbf{y}\|_1]], \quad (11)$$

where G is the generator mapping from domain X to domain Y , F is the generator mapping from domain Y to domain X , $\mathbf{x} \in X$ is an image from domain X , and $\mathbf{y} \in Y$ is an image from domain Y .

For sensor simulation, these models are typically employed for tasks such as sim-to-real mapping or weather modelling, where creating paired data in different domains is super challenging. The summary of unpaired I2I translation models based on GANs is provided in Table IV. In the following, we will separately review the unpaired I2I models for synthesising Camera RGB images and Lidar point clouds.

1) *RGB Image synthesis*: Building on CycleGAN's concept, most unpaired I2I models have introduced methods to enforce consistency. DistanceGAN [39] maintains structural relationships without inverse mapping, effective for translating different weather conditions. CyCADA [40] combines pixel-level and feature-level domain adaptation with cycle-consistency constraints, preserving semantic content. MUNIT [41] uses content and style codes to improve diversity and quality. The CUT model [42] employs contrastive learning to enhance translation by maximising mutual information, and simplifying training. SRUNIT [43] introduces a robustness loss for semantic invariance, improving semantic integrity. Spatially-correlative loss [44] preserves scene structure while allowing appearance changes, enhancing structural consistency. Jung et al. [45] use semantic relation consistency regularisation and decoupled contrastive learning to improve spatial correspondence and semantic alignment. SPR [46] finds the shortest path between domains, enhancing content preservation while transforming appearance.

The other models in this domain are miscellaneous, as reviewed in the following paragraphs.

Sensor Transfer [51] tackles the domain gap between synthetic and real datasets by transferring sensor effects such as chromatic aberration, blur, exposure, noise, and colour temperature from real datasets to synthetic ones. This approach improves the robustness and generalisability of models trained on synthetic data when applied to real-world tasks, enhancing

TABLE IV: Summary of unpaired I2I translation GAN-based models.

Model	Sensor	Year	Description	Datasets
CycleGAN [38]	Camera	2017	Introduces cycle-consistency loss for unpaired image-to-image translation	Cityscapes [21]
DistanceGAN [39]	Camera	2017	Maintains structural relationships within images during translation	Cityscapes [21]
CyCADA [40]	Camera	2017	Combines pixel-level and feature-level domain adaptation with cycle-consistency constraints	GTA-V [47], SYNTHIA [48]
MUNIT [41]	Camera	2018	Decomposes image representation into content and style codes for diverse outputs	Cityscapes [21], SYNTHIA [48]
Saleh et al. [49]	Lidar	2019	Employs CycleGAN for sim-to-real mapping on BEV Lidar point clouds	KITTI [19], CARLA [50]
Sensor Transfer [51]	Camera	2018	Transfers sensor effects from real datasets to synthetic ones to improve robustness	KITTI [19], GTA-V [47], Cityscapes [21]
CUT [42]	Camera	2020	Uses contrastive learning to enhance unpaired I2I translation	Cityscapes [21]
AnalogicalGAN [52]	Camera	2020	Learns from synthetic images to perform zero-shot image translation for foggy scenes	Cityscapes [21], V-KITTI [53]
Tremblay et al. [54]	Camera	2020	Integrates multiple methods for realistic rain augmentation on image datasets	KITTI [19], Cityscapes [21], nuScenes [55]
SRUNIT [43]	Camera	2021	Enforces semantic robustness to address semantics flipping in image translation	Cityscapes [21]
Richter et al. [56]	Camera	2021	Enhances realism of synthetic images by leveraging intermediate representations	GTA-V [47], Cityscapes [21]
PCT [57]	Lidar	2021	Decomposes synthetic-to-real gap into appearance and sparsity components	SynLiDAR, Semantic-KITTI [58]
USIS [59]	Camera	2021	Uses SPADE generator with self-supervised segmentation loss for realistic image synthesis	Cityscapes [21]
Jung et al. [45]	Camera	2022	Enhances spatial correspondence and semantic alignment with semantic relation consistency	Cityscapes [21]
Eskandar et al. [60]	Camera	2023	Uses synthetic semantic layout to generate real RGB images maintaining content integrity	GTA-V [47], Cityscapes [21], Mapillary [61]
SPR [46]	Camera	2023	Encourages network to find shortest path connecting two domains in unpaired translation	Cityscapes [21]
Barrera et al. [62]	Lidar	2023	Preserves small object details during domain adaptation on BEV point clouds	CARLA [50], KITTI [19]
CLS2R [63]	Lidar	2023	Uses contrastive learning for sim-to-real mapping of Lidar point clouds	CARLA [50], KITTI [19]

the realism of synthetic data more effectively than traditional domain randomisation techniques.

Richter et al. [56] enhance the realism of synthetic images by leveraging intermediate representations produced by conventional rendering pipelines. Their method integrates a convolutional network trained via a novel adversarial objective, aligning scene layout distributions across datasets to reduce artefacts and significantly enhance the stability and realism of synthesised images. The USIS [59] model marks a significant advance in creating realistic images from segmentation masks without paired data. This framework leverages a SPADE generator enhanced with a self-supervised segmentation loss and a wavelet-based discriminator, ensuring high semantic consistency and detailed textures.

Eskandar et al. [60] propose a framework for pragmatic semantic image synthesis (SIS) for urban scenes, using synthetic semantic layout to generate real RGB images that maintain the content of the input mask while adopting the appearance of real images.

AnalogicalGAN [52] introduces the concept of analogical image translation, learning from synthetic clear-weather and foggy images to translate real clear-weather images to real foggy images without seeing any real foggy images during training. This approach allows AnalogicalGAN to perform zero-shot image translation [53].

Tremblay et al. [54] present a comprehensive rain rendering pipeline designed to improve the robustness of computer vision algorithms under rainy conditions. This approach integrates physics-based, data-driven, and hybrid methods to generate realistic rain effects on existing image datasets. The importance of realistic rain augmentation is highlighted through extensive validation on datasets such as KITTI, Cityscapes, and nuScenes [55].

2) *Lidar Point Cloud Synthesis*: Unpaired I2I models for Lidar point cloud synthesis usually first transform the 3D point cloud into range images which are further processed by image-based models. The first notable model [49] employs a CycleGAN-based framework to address the domain shift between synthetic and real Lidar point cloud data, particularly for vehicle detection from a bird’s eye view (BEV). This model enhances previous methods by maintaining the structural integrity of real Lidar data during translation, significantly improving vehicle detection performance.

The ePointDA model [67] presents an end-to-end sim-to-real domain adaptation framework for Lidar point cloud segmentation, bridging the domain gap at both pixel and feature levels. Unlike previous multi-stage pipelines, it includes self-supervised dropout noise rendering, statistics-invariant and spatially-adaptive feature alignment, and transferable segmentation learning. This approach enhances Lidar segmentation

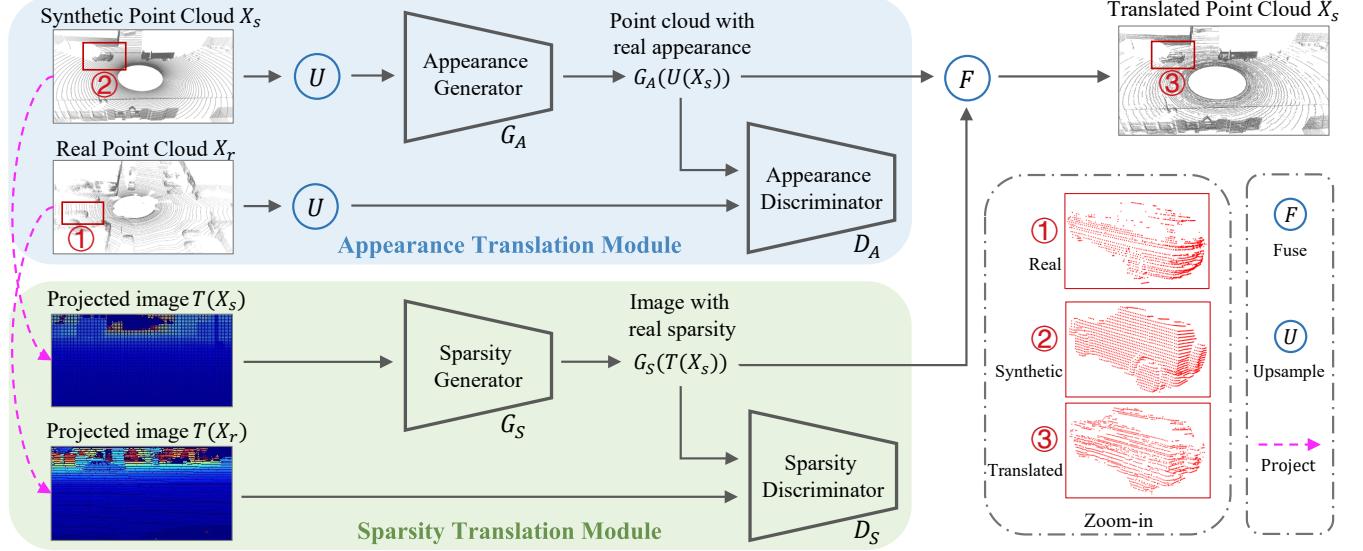


Fig. 5: The training data flow of PCT [57], an unpaired data translation model based on GANs for sim-to-real mapping of Lidar point clouds.

TABLE V: Summary of GAN-based video Prediction models (all used for camera RGB image synthesis).

Model	Year	Description	Datasets
GeoSim [64]	2021	Creates realistic video simulations using 3D objects and a physics engine.	Argoverse [65]
DriveGAN [66]	2021	Learns from video footage and actions for interactive scene editing and future frame prediction.	CARLA [50], RWD [66]

by rendering dropout noise for synthetic data and aligning features spatially between domains without relying on real-world statistics.

The Barrera et al. [62] build upon the CycleGAN framework by incorporating semantic consistency to preserve small object details during domain adaptation on BEV point clouds. This model excels in maintaining information about small objects, such as pedestrians and cyclists, which are often lost in traditional methods.

The PCT (Point Cloud Translator) model [57] addresses domain gaps between synthetic and real Lidar point clouds by decomposing the synthetic-to-real gap into appearance and sparsity components. It employs an Appearance Translation Module (ATM) to up-sample synthetic point clouds and translate them to resemble real point clouds, followed by a Sparsity Translation Module (STM) to integrate real sparsity features (see Fig. 5). This innovative approach effectively mitigates domain gaps, resulting in high-quality translations that enhance semantic segmentation tasks.

The CLS2R model [63] introduces a contrastive learning framework for sim-to-real mapping of Lidar point clouds, using a lossless representation of Lidar data that includes depth, reflectance, and raydrop attributes. This model enhances realism and faithfulness by using contrastive learning to ensure high similarity between input and output patches, effectively synthesising realistic Lidar point clouds.

D. Video Prediction Models

Video prediction models represent another type of conditional model, where the goal is to predict future video frames

conditioned on the preceding ones. In opposition to previous GAN categories, these models incorporate the temporal element, requiring them to effectively model the movement of dynamic objects. The summary of video prediction models based on GANs is provided in Table V. In the following, we review two notable works in this area, both applied to RGB image synthesis.

GeoSim [64] introduces a geometry-aware image composition process for creating realistic video simulations for self-driving applications. This model tackles the challenges of photo-realism and high-level control by utilising a diverse bank of 3D objects derived from sensor data. By proposing realistic object placements, rendering dynamic objects in new poses, and seamlessly blending them into existing scenes, GeoSim ensures traffic-aware, geometrically consistent synthetic images. Furthermore, the model incorporates a physics engine to simulate realistic vehicle dynamics and interactions, significantly enhancing the overall realism of the simulations.

DriveGAN [66] employs a neural simulator that learns from sequences of video footage and the actions taken by an ego-agent within an environment. Utilising a VAE and GANs, it learns a latent space for images, enabling the dynamics engine to learn transitions within this space (as shown in Fig. 6). DriveGAN distinguishes itself by disentangling different components of a scene without supervision, allowing users to interactively edit scenes and generate unique scenarios. Additionally, the model features an action-conditioned component that predicts future frames based on the current state and intended actions, offering a robust framework for simulating driving behaviours.

TABLE VI: Summary of unconditional diffusion models.

Model	Sensor	Year	Description	Datasets
LiDARGen [68]	Lidar	2022	Generates point clouds using a stochastic denoising process ensuring physical feasibility.	KITTI-360 [69], NuScenes [55]
Parke et al. [70]	Camera	2023	Simultaneously generate images and semantic layouts with a combined diffusion model.	Cityscapes [21]
R2DM [71]	Lidar	2024	Uses a diffusion model to generate diverse and high-fidelity 3D scene point clouds.	KITTI-360 [69], KITTI-Raw [19]
RangeLDM [72]	Lidar	2024	Utilises latent diffusion to generate high-quality range-view Lidar point clouds.	KITTI-360 [69], NuScenes [55]
LiDM [73]	Lidar	2024	State-of-the-art method for generating realistic Lidar scenes with advanced techniques.	KITTI-360 [69], NuScenes [55]

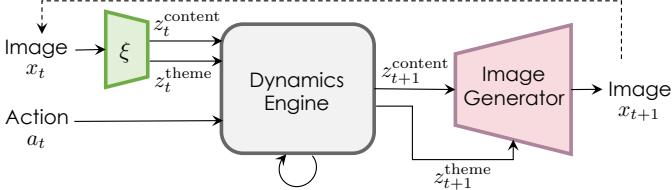


Fig. 6: The inference pipeline of DriveGAN [66], a video prediction model based on GANs for RGB images.

IV. GENERATIVE MODELS – DIFFUSION MODELS

Diffusion models have set new benchmarks in various data generation tasks due to their inherent advantages over GAN models, such as more stable training and the ability to iteratively refine samples, resulting in higher quality synthesis. These models have found extensive applications in sensor simulation for autonomous driving, spanning a range of settings from basic unconditional models to text-to-image models and even spatiotemporal video prediction models, as discussed below.

A. Unconditional Models

Unconditional diffusion models generate data from a simple prior distribution, such as Gaussian noise, and iteratively refine it through a series of transformations driven by the diffusion process (as discussed in Section II). The summary of unconditional DMs is provided in Table VI. The following sections separately review the unconditional diffusion models for both camera image synthesis and Lidar point cloud generation.

1) *RGB Image Synthesis*: In the area of unconditional diffusion models for RGB images, we identified a single notable work. The Gaussian-categorical diffusion process [70] model presents an RGB image synthesis model by simultaneously generating images and their corresponding semantic layouts. This technique enhances image quality by incorporating semantic understanding into the generation process, using a combined Gaussian and categorical diffusion model to represent the joint distribution of image-layout pairs.

2) *Lidar Point Cloud Synthesis*: The pioneering approach that uses diffusion models for unconditional Lidar point cloud generation is LiDARGen [68]. This model formulates the point cloud generation process as a stochastic denoising process in the equirectangular view. LiDARGen advances previous methods by ensuring the physical feasibility and controllability

of generated samples. Its capability to sample point clouds conditioned on inputs without retraining makes LiDARGen particularly useful for Lidar densification tasks in ADS.

Building on this, the R2DM [71] introduces a DM-based framework for Lidar data. R2DM generates diverse and high-fidelity 3D scene point clouds based on the image representation of range and reflectance intensity. This model focuses on stable training, sample quality, and versatility in handling inverse problems. The R2DM model has demonstrated remarkable performance on the KITTI-360 and KITTI-Raw datasets, particularly excelling in Lidar completion tasks.

RangeLDM [72] leverages latent diffusion models to rapidly generate high-quality range-view Lidar point clouds. This model projects point clouds onto range images using Hough Voting for accurate range-view data distribution. The range images are then compressed into a latent space with a VAE and processed by a diffusion model to enhance expressivity.

Finally, the LiDM [73] model presents a SOTA approach for generating realistic Lidar scenes using diffusion models. LiDM focuses on pattern realism, geometry realism, and object realism by introducing three core innovations: curve-wise compression, point-wise coordinate supervision, and patch-wise encoding. These techniques ensure that the generated Lidar scenes maintain high fidelity to real-world data, preserving the detailed geometric and structural properties of objects and their surroundings.

B. Text-to-Image Translation Models

Building on unconditional DMs, several methods have incorporated auxiliary data for enhanced control and synthesis quality. For instance, the Semantic Diffusion Model (SDM) [74] processes semantic layouts and noisy images separately, embedding the semantic layout into the decoder of the denoising network. The reverse diffusion process for this model can be described as:

$$q(\mathbf{y}_{t-1} | \mathbf{y}_t, \mathbf{x}) = \mathcal{N}(\mathbf{y}_{t-1}; \mu_\theta(\mathbf{y}_t, t, \mathbf{x}), \Sigma_\theta(\mathbf{y}_t, t, \mathbf{x})). \quad (12)$$

where \mathbf{x} represents the condition, such as a semantic layout or text prompt, and \mathbf{y} is the image being modelled. The summary of T2I translation DMs is provided in Table VII. The following sections will review other conditional diffusion models, all of which have been applied to RGB image synthesis.

Expanding on SDM's advancements, GEODIFFUSION [75] proposes a text-prompted geometric control framework for

TABLE VII: Summary of T2I diffusion models (all used for camera RGB image synthesis).

Model	Year	Description	Datasets
SDM [74]	2022	Processes semantic layouts and noisy images separately, embedding the layout into the denoising network.	Cityscapes [21]
GEODIFFUSION [75]	2023	Proposes a text-prompted geometric control framework for high-quality object detection data.	NuScenes [55]
BEVControl [76]	2023	Generates realistic and controllable street-view images from BEV sketches.	NuScenes [55]
MAGICDRIVE [77]	2023	Uses 3D geometry controls like camera poses, road maps, and bounding boxes with textual descriptions.	NuScenes [55]
Loiseau et al. [78]	2023	Generates synthetic data to assess perception model reliability under domain shifts.	Cityscapes [21]
Text2Street [79]	2024	Framework for generating controllable street-view images from text, including road topology and weather.	NuScenes [55]

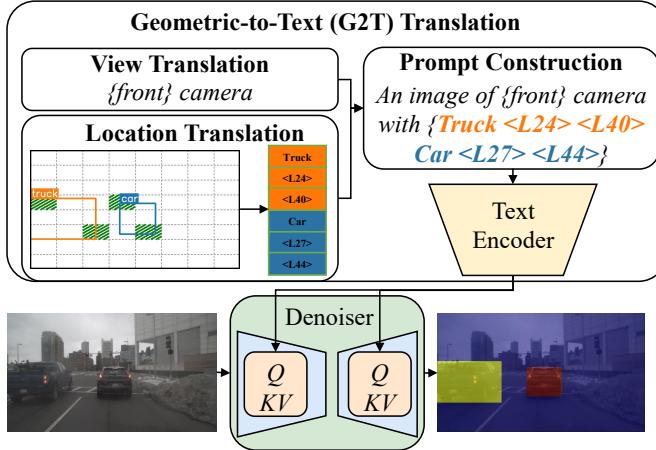


Fig. 7: The model architecture of GeoDiffusion [75], a T2I translation diffusion model for RGB images.

generating high-quality object detection data. This method leverages pre-trained T2I diffusion models to translate various geometric conditions into text prompts, as shown in Fig. 7. In contrast to previous layout-to-image methods that encoded only bounding boxes, GEODIFFUSION incorporates additional geometric conditions such as camera views, enhancing the realism and consistency of generated images.

Further refining this concept, BEVControl [76] introduces a robust approach for generating realistic and controllable street-view images from BEV sketches. By incorporating a two-stage framework that decouples visual consistency into geometry and appearance sub-goals, BEVControl allows for detailed manipulation of background and foreground elements.

Taking this approach further, MAGICDRIVE [77] incorporates diverse 3D geometry controls, such as camera poses, road maps, and 3D bounding boxes, along with textual descriptions. By using a cross-view attention module, MAGICDRIVE ensures consistency across multiple camera views, significantly enhancing the realism and geometric accuracy of generated scenes. Its ability to control various scene attributes, such as weather conditions and time of day, is particularly useful for BEV segmentation and 3D object detection tasks.

Loiseau et al. [78] introduce an innovative approach to generating realistic synthetic data for assessing the reliability of perception models under various domain shifts. By leveraging a pre-trained T2I diffusion model, specifically Stable

Diffusion [89], augmented with a ControlNet [90] module, this method conditions generation on semantic masks from the Cityscapes dataset while using text prompts to simulate different target domains. This approach enables zero-shot generation of synthetic data that aligns with the semantic conditions of the original domain while reflecting the visual properties of the target domains.

Finally, Text2Street [79] offers a framework for generating controllable street-view images from textual descriptions. Introducing three main components—a lane-aware road topology generator, a position-based object layout generator, and a multiple control image generator—Text2Street generates detailed and semantically accurate street-view images based on specific textual descriptions of road topology, traffic status, and weather conditions.

C. Video Prediction Models

Video prediction diffusion models (also referred to as world models) are a form of conditional diffusion models where the input consists of previous frames and the goal is to predict future ones. These models differ from standard conditional diffusion models by needing to model temporal interactions and being more computationally efficient due to higher input dimensionality. They can also incorporate various auxiliary conditions, such as actions, bounding box layouts, and text prompts, to enhance controllability. The summary of video prediction DMs is provided in Table VIII. The following sections review several pioneering works in this domain for both RGB image and Lidar point cloud synthesis.

1) *RGB Image Synthesis*: DriveDreamer [80] is a pioneering model that constructs comprehensive world models from real-world driving videos and human driver behaviours. It introduces the Autonomous-driving Diffusion Model (Auto-DM) with a two-stage training pipeline, as shown in Fig. 8. The first stage enhances sampling efficiency by incorporating traffic structural information, while the second stage focuses on video prediction to establish the world model. This approach generates high-quality, controllable driving videos and reasonable driving policies.

Building on this foundation, Drive-WM [81] is the first driving world model compatible with existing end-to-end planning models. It features joint spatial-temporal modelling facilitated by view factorisation, enabling the generation of high-fidelity multi-view videos in driving scenes. Drive-WM

TABLE VIII: Summary of diffusion video prediction models.

Model	Sensor	Year	Description	Datasets
DriveDreamer [80]	Camera	2023	Constructs world models from driving videos and driver behaviours, enhancing sampling efficiency with traffic structural information.	NuScenes [55]
Drive-WM [81]	Camera	2023	Features joint spatial-temporal modelling for generating high-fidelity multi-view videos in driving scenes.	NuScenes [55]
DrivingDiffusion [82]	Camera	2023	Introduces a spatial-temporal consistent diffusion framework for generating realistic multi-view videos controlled by 3D layout.	NuScenes [55]
Panacea [83]	Camera	2023	Generates panoramic and controllable videos for autonomous driving scenarios with a two-stage system and 4D attention mechanism.	NuScenes [55]
ADriver-I [84]	Camera	2023	Unifies control signal prediction and future scene generation within a single framework using a multimodal large language model.	NuScenes [55]
WoVoGen [85]	Camera	2023	Combines an explicit 4D world volume with diffusion models to generate high-quality, multi-camera street-view videos.	NuScenes [55]
SubjectDrive [86]	Camera	2024	Enhances the scalability and diversity of generative data by integrating external subjects into the generation process.	NuScenes [55]
DriveDreamer-2 [87]	Camera	2024	Enhances DriveDreamer with a Large Language Model to generate user-defined driving videos from text prompts.	NuScenes [55]
LidarDM [88]	Lidar	2024	Produces realistic, layout-aware, and temporally coherent Lidar point cloud videos, useful for autonomous driving simulations.	KITTI [19], NuScenes [55]

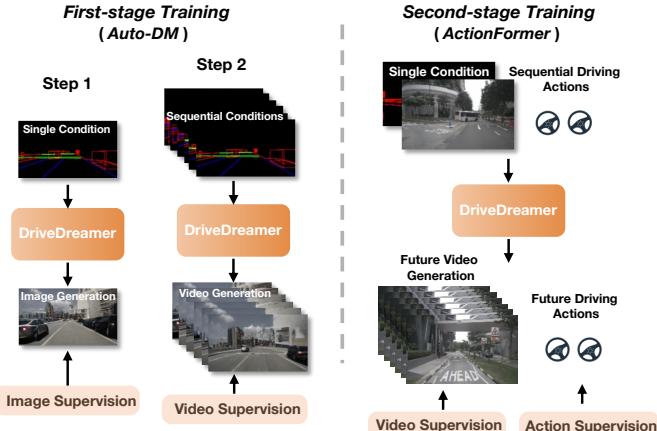


Fig. 8: The two-stage training pipeline of DriveDreamer [80], a video prediction diffusion model for RGB images.

predicts multiple futures based on distinct driving manoeuvres and determines the optimal trajectory using image-based rewards. Drive-WM leverages latent video diffusion models for multi-view and temporal modelling, ensuring consistency and quality across different views and frames.

DrivingDiffusion [82] introduces a spatial-temporal consistent diffusion framework for generating realistic multi-view videos controlled by 3D layout. The method addresses three main challenges: cross-view consistency, cross-frame consistency, and instance quality. The multi-stage scheme includes a multi-view single-frame image generation model, a single-view video generation step shared by multiple cameras, and post-processing to enhance video length and consistency. The use of information exchange between adjacent cameras and a temporal sliding window algorithm ensures the generation of high-quality, realistic driving videos. Panacea [83] introduces a novel approach for generating panoramic and controllable videos tailored for autonomous driving scenarios. This method

addresses key challenges in video generation: ensuring consistency and controllability. Panacea operates through a two-stage system where the first stage generates multi-view driving scene images, and the second stage extends these images temporally to create video sequences. It incorporates a novel 4D attention mechanism and the ControlNet framework for detailed control using BEV layouts.

ADriver-I [84] proposes an innovative method that unifies control signal prediction and future scene generation within a single framework. It uses interleaved vision-action pairs as inputs and employs a multimodal large language model (MLLM) and video DM (VDM) to iteratively predict control signals and future frames. Unlike previous methods, ADriver-I does not require extensive prior information such as 3D bounding boxes or HD maps, making it more flexible and generalisable.

WoVoGen [85] combines an explicit 4D world volume with diffusion models to generate high-quality, multi-camera street-view videos. This model operates in two phases: envisioning a 4D temporal world volume based on vehicle control sequences and generating multi-camera videos informed by this envisioned world volume. By incorporating a dense voxel volume that encapsulates comprehensive data about the scene, WoVoGen ensures both intra-world consistency and inter-sensor coherence.

SubjectDrive [86] presents an innovative video generation framework aimed at enhancing the scalability and diversity of generative data. It integrates external subjects into the generation process via a subject control mechanism, significantly boosting data diversity and utility. The architecture comprises three modules: the Subject Prompt Adapter (SPA) for enriching text embeddings, the Subject Visual Adapter (SVA) for incorporating spatial information, and Augmented Temporal Attention (ATA) for ensuring temporal consistency.

Building on this progress, DriveDreamer-2 [87] enhances the DriveDreamer framework by incorporating an LLM to generate user-defined driving videos from text prompts. The

TABLE IX: Summary of miscellaneous generative models.

Model	Category	Year	Description	Datasets
CRN [91]	Supervised I2I	2017	Synthesises photographic images from semantic layouts using cascaded refinement modules, without adversarial training.	Cityscapes [21]
Vecek et al. [92]	Supervised I2I	2020	Predicts the strength of Lidar responses using deep learning, enhancing data realism and improving segmentation accuracy.	GTA V [47], semantic-KITTI [58]
LiDARsim [93]	Supervised I2I	2020	Generates realistic Lidar point clouds by leveraging real-world data, combining physics-based simulation and deep neural network refinement.	KITTI [19], NuScenes [55]
Hu et al. [94]	AR Video Prediction	2020	Predicts ego-motion, static scenes, and dynamic agent motion probabilistically using a spatio-temporal convolutional module.	Cityscapes [21]
RINet [95]	Supervised I2I	2022	Simulates Lidar sensors by mapping RGB images to corresponding Lidar features, significantly improving Lidar data realism.	CARLA [50], Waymo [33], semantic-KITTI [58]
READ [96]	Supervised I2I	2022	Uses neural rendering to generate realistic driving scenes from sparse point clouds with multi-sampling for coherence and efficiency.	KITTI [19], Brno Urban [97]
UltraLiDAR [98]	Unconditional AR	2023	Uses a VQ-VAE framework to learn compact, discrete 3D representations of scene-level Lidar point clouds, enabling various downstream applications.	Pandaset [99], KITTI-360 [69]
MUVO [100]	AR Video Prediction	2023	Integrates raw camera and Lidar data to learn a sensor-agnostic geometric world representation, enhancing prediction quality.	CARLA [50]
ETSSR [101]	Supervised I2I	2023	Accelerates stereo image simulation using Stereo Super Resolution (SSR) with a transformer-based model inspired by Swin Transformer.	CARLA [50]
WorldDreamer [102]	AR Video Prediction	2024	Uses Spatial Temporal Patchwise Transformer (STPT) for visual token prediction, integrating multi-modal prompts for video generation.	NuScenes [55]
LidarGRIT [103]	Unconditional AR	2024	Addresses limitations in generating realistic raydrop noise with a generative range image transformer model using AR transformer and VQ-VAE.	KITTI-360 [69], KITTI odometry [19]
BEVGen [104]	Supervised I2I	2024	Synthesises spatially consistent street-view images conditioned on BEV layouts using VQ-VAE auto-encoders and a causal transformer.	nuScenes [55], Argoverse 2 [105]

method separates traffic simulation into foreground (agent trajectories) and background (HDMaps) conditions, using a functional library to finetune the LLM for trajectory generation. DriveDreamer-2 introduces the Unified Multi-View Model (UniMVM) to ensure temporal and spatial coherence in the generated videos, significantly improving video quality.

2) *Lidar Point Cloud Synthesis*: In the field of Lidar point cloud video prediction, a notable approach is LidarDM [88]. This model introduces a generative method capable of producing realistic, layout-aware, and temporally coherent Lidar point cloud videos. In contrast to previous models, LidarDM incorporates driving scenario guidance, making it particularly useful for autonomous driving simulations. It uses a conditional generative framework to ensure the generated Lidar data is physically plausible and aligns with the expected layout of driving environments.

V. GENERATIVE MODELS – MISCELLANEOUS

Other generative models used for camera and Lidar simulation are often disintegrated into distinct categories. The summary of miscellaneous generative models is provided in Table IX. In the following, we review these models in three sub-categories: unconditional AR models, supervised I2I translation models, and AR video prediction models.

A. Unconditional Auto-regressive Models

Unconditional AR models generate data sequentially without relying on any conditional input, predicting the next

value in a sequence based on previously generated values, as discussed in Section II. Two notable models in this category, used for Lidar point cloud generation, are UltraLiDAR [98] and LidarGRIT [103].

UltraLiDAR [98] builds upon the concept of learning compact, discrete 3D representations of scene-level Lidar point clouds, enabling various downstream applications including unconditional Lidar generation. UltraLiDAR focuses on learning a discrete codebook that captures the geometric structure of scenes and aligns sparse point clouds with dense ones. This model employs a VQ-VAE [106] framework and introduces techniques to handle voxelised BEV representation, enhancing computational efficiency.

LidarGRIT [103] introduces the Lidar generative range image transformer model, which addresses previous DMs' limitations in generating realistic raydrop noise. LidarGRIT uses progressive generation and accurate raydrop noise synthesis through iterative sampling in the latent space via an AR transformer. The sampled tokens are then decoded to range images using an adapted VQ-VAE (see Fig. 9). By separating the generation of range images from raydrop noise masks, this method significantly enhances data realism.

B. Supervised Image-to-Image Translation Models

Several miscellaneous models can be classified as supervised I2I models. These models do not rely on adversarial, denoising diffusion, or AR prediction losses; instead, they primarily leverage supervised error losses such as perceptual

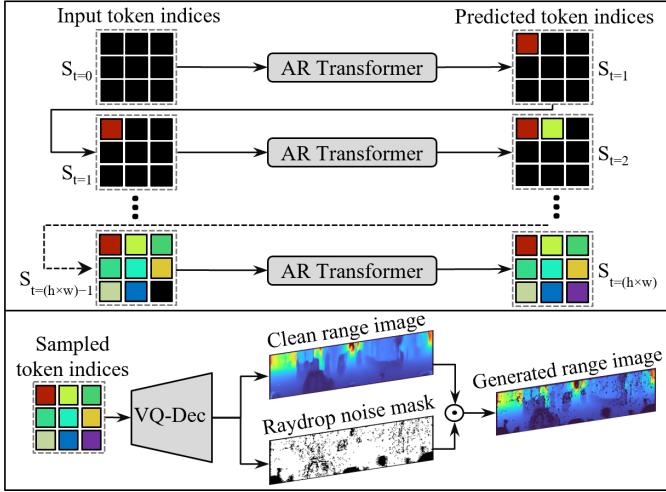


Fig. 9: Inference process of the unconditional auto-regressive model, LidarGRIT [103], for Lidar point cloud generation.

losses or L1/L2 error losses. In the following sections, we review these models, categorising them into either Camera RGB image synthesis or Lidar point cloud synthesis.

1) *RGB Image Synthesis*: CRN model [91] introduces a feedforward network approach to synthesise photographic images from semantic layouts without the need for adversarial training. This model uses cascaded refinement modules to process input semantic layouts through multiple stages, progressively enhancing the resolution and detail of the output image. Each module refines the image by combining downsampled input layouts with upsampled features from the previous module.

READ model [96] introduces a novel neural rendering approach to synthesise photo-realistic driving scenes from sparse point clouds. This method leverages a neural network, which filters and fuses features across different scales to produce detailed and coherent images. In contrast to previous models that struggled with coherence and computational efficiency, READ employs multi-sampling strategies such as Monte Carlo sampling and patch sampling to optimise performance and reduce computational costs. ETSSR model [101] presents a novel method to accelerate stereo image simulation by employing stereo super resolution. The proposed technique initially simulates low-resolution stereo images, then super-resolves them to high-resolution using a novel transformer-based SSR model inspired by Swin Transformer [107]. This method leverages a Disparity-aware Swin Cross Attention Module (DSCAM) for efficient cross-view feature extraction, enhancing both the realism and computational efficiency of the generated images.

BEVGen model [104] employs a novel generative model to synthesise spatially consistent street-view images conditioned on BEV layouts. The model integrates two VQ-VAE auto-encoders for images and BEV representations, and a causal transformer to model high-level scenes. The unique aspect of BEVGen is its ability to handle cross-modal inductive 3D bias, which enables the model to relate information between modalities and across different views. This approach achieves high-

quality synthesis results and offers practical applications such as data augmentation and simulated driving scene rendering.

2) *Lidar Point Cloud Synthesis*: Vacek et al. [92] introduce a data-driven method for simulating Lidar sensors by predicting the strength of Lidar responses. This model uses computer-generated data to extract geometrically simulated point clouds and predicts Lidar intensities using deep learning. LPI enhances data realism by accounting for systematic failures and noise, such as low responses on polished surfaces and strong responses on reflective surfaces.

LiDARsim [93] presents a novel approach for generating realistic Lidar point clouds by leveraging real-world data. This method involves creating a large catalogue of 3D static maps and dynamic objects by driving around several cities with a self-driving fleet. These assets are then used to compose scenes where a physics-based simulator first performs ray casting over the 3D scene, and a deep neural network refines the output to produce realistic Lidar point clouds. This hybrid approach of combining physics-based and learning-based simulations allows LiDARsim to capture more complex interactions and sensor noise.

Guillard et al. [95] introduce a data-driven pipeline to simulate Lidar sensors by mapping RGB images to corresponding Lidar features, such as raydrop and per-point intensities, using real datasets. The Raydrop and Intensity Network (RINet) predicts realistic Lidar characteristics from RGB images, enhancing ray-casted point clouds from standard simulation software (see Fig. 10). This approach encodes effects such as dropped points on transparent surfaces and high-intensity returns on reflective materials, significantly improving Lidar data realism.

C. Auto-regressive Video Prediction Models

There are several video prediction models (i.e. world models) that are primarily based on auto-regressive approaches, utilising spatio-temporal modelling with RNN/GRU models or causal transformers. In the following sections, we will review papers in this domain, all of which have been applied to camera RGB image synthesis or joint RGB image and Lidar point cloud synthesis.

Hu et al. [94] introduce a deep learning architecture designed for probabilistic future prediction from video data. This model is the first to jointly predict ego-motion, static scenes, and dynamic agent motion in a probabilistic manner. It leverages a spatio-temporal convolutional module, including ConvGRU, to learn representations from RGB video, which can be decoded into future semantic segmentation, depth, and optical flow. The model employs a conditional variational approach to minimise the divergence between present and future distributions, enabling the generation of diverse and accurate future scenarios.

MUVO [100] introduces a multi-modal generative world model with geometric voxel representations, integrating raw camera and Lidar data to learn a sensor-agnostic geometric world representation. In contrast to previous models that focused solely on sensor data, MUVO enhances prediction quality by incorporating 3D occupancy predictions conditioned

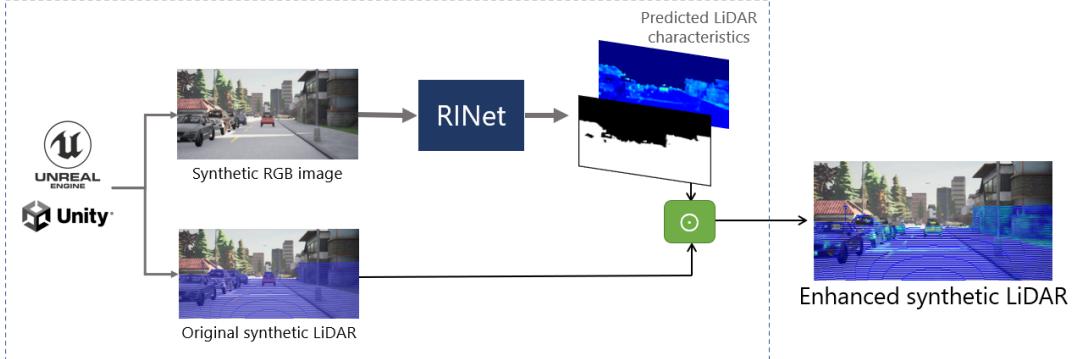


Fig. 10: Inference pipeline of RINet [95], a supervised I2I model for Lidar raydrop synthesis.

TABLE X: Summary of frequency-based NeRFs (all used for camera RGB image synthesis).

Model	Year	Description	Datasets
NSG [108]	2020	Decomposes dynamic scenes into scene graphs, representing each object as nodes within a hierarchical structure, enabling efficient rendering of dynamic scenes.	KITTI [19]
Block-NeRF [109]	2022	Divides the environment into multiple compact NeRFs, each trained independently, allowing efficient updates without retraining the entire model.	Alamo Square [109]
MapNeRF [110]	2023	Incorporates map priors into neural radiance fields to synthesise out-of-trajectory driving views with semantic road consistency.	Argoverse2 [105]
UC-NeRF [111]	2023	Introduces layer-based colour correction, virtual warping, and spatiotemporally constrained pose refinement for high-quality neural rendering with multiple cameras.	Waymo [33], NuScenes [55]
ChatSim [112]	2024	Enables editable photo-realistic 3D driving scene simulations via natural language commands with external digital assets, using McNeRF and McLight methods.	Waymo [33]

on actions, showing improvements in both camera and Lidar data prediction.

WorldDreamer [102] advances the concept of world models for video generation by framing the task as a visual token prediction challenge. This model utilises the Spatial-Temporal Patchwise Transformer (STPT) to focus attention on localised patches within a temporal-spatial window. WorldDreamer excels in generating videos across different scenarios, including natural scenes and driving environments.

VI. VOLUME RENDERERS – NERFS

Neural Radiance Fields (NeRFs) [10] have recently gained significant attention for scene representation and sensor simulation in ADS due to their exceptional ability to produce high-quality novel view synthesis. In contrast to the black-box nature of generative models, NeRFs provide a more transparent and explicit modelling of scenes and radiance. As discussed in Section VI, NeRFs use MLP to model the colour and density of a 3D position $\mathbf{x} \in \mathbb{R}^3$ and the direction of a 2D ray $\mathbf{d} \in \mathbb{R}^2$. The encoding of this 5D input is crucial for the performance of NeRFs. In the context of ADS, various encoding methods, such as frequency-based, hash-grids, point-based, and multi-planar, are utilised, as discussed in the following sections.

A. Frequency-based Encoding

Frequency-based encoding is the first technique used for positional encoding and has proven highly effective in capturing and rendering high-frequency details in 3D scenes. This

encoding transforms the input 3D coordinates \mathbf{x} and viewing direction \mathbf{d} into a higher-dimensional space using high-frequency functions. The frequency-based encoding function γ can be defined as follows:

$$\begin{aligned} \gamma(\mathbf{p}) = & (\sin(2^0\pi\mathbf{p}), \cos(2^0\pi\mathbf{p}), \\ & \sin(2^1\pi\mathbf{p}), \cos(2^1\pi\mathbf{p}), \dots, \sin(2^{L-1}\pi\mathbf{p}), \cos(2^{L-1}\pi\mathbf{p})), \end{aligned} \quad (13)$$

where $\mathbf{p} \in \mathbb{R}^3$ can be either the 3D position \mathbf{x} or the 2D viewing direction \mathbf{d} , and L is the number of frequency bands used in the encoding. The summary of frequency-based NeRFs is provided in Table X. In the following, we review NeRFs based on frequency-based encoding, which are all used for RGB image synthesis in ADS.

NSG model [108] introduces a novel neural rendering method that decomposes dynamic scenes into scene graphs, representing each object as nodes within a hierarchical structure. This approach enables efficient and accurate rendering of dynamic scenes by leveraging individual neural radiance fields for each object, in contrast to previous methods that encoded the entire scene into a single network.

Block-NeRF [109] extends NeRFs to large-scale scenes by dividing the environment into multiple compact NeRFs, each trained independently to cover a specific block of the environment (see Fig. 11). This method addresses the scalability issue of NeRFs, allowing for efficient updates to individual blocks without retraining the entire model. Key innovations include using appearance embeddings, learned pose refinement and controllable exposure to handle data captured under varying environmental conditions.

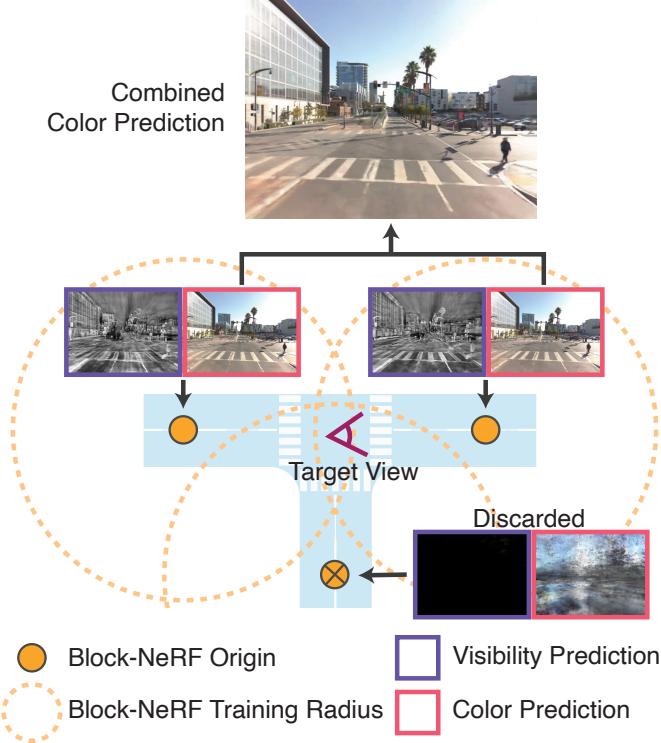


Fig. 11: Scene split in Block-NeRF [109], a volume renderer using frequency-based NeRFs for RGB image synthesis.

Further enhancing the concept of scene representation, MapNeRF [110] introduces the incorporation of map priors into neural radiance fields to synthesise out-of-trajectory driving views with semantic road consistency. This method leverages ground and lane information from maps to supervise the density field and warp depth, ensuring multi-view consistency even when the camera deviates from the standard trajectory.

Addressing the challenges in multi-camera systems, UC-NeRF [111] introduces a method for high-quality neural rendering with multiple cameras in autonomous driving. The key contributions include layer-based colour correction, virtual warping to generate diverse but consistent virtual views, and spatiotemporally constrained pose refinement for robust pose calibration. This approach significantly improves the rendering quality and accuracy of synthesised novel views.

Finally, ChatSim [112] presents a novel system enabling editable photo-realistic 3D driving scene simulations via natural language commands with external digital assets. This method leverages a collaborative framework of LLM agents to decouple complex simulation tasks into specific editing operations. The key contributions include McNeRF, a multi-camera neural radiance field method, and McLight, a multi-camera lighting estimation method for the scene-consistent rendering of external digital assets.

B. Hash-Grids

Hash-grid encoding [129] is the largest subcategory of NeRFs as it is the most effective and computationally efficient method for scene representation. This encoding technique

utilises a multi-resolution hash table to map input coordinates to feature vectors efficiently. In hash-grid encoding, a spatial point $\mathbf{p} \in \mathbb{R}^3$ is transformed by hashing its coordinates at multiple resolutions, enabling the network to capture both fine and coarse details. The transformation is defined as follows:

$$\mathbf{h}_i(\mathbf{p}) = \text{hash}(\mathbf{p} \cdot \mathbf{R}_i) \bmod T, \quad (14)$$

where \mathbf{h}_i is the hashed index for the i^{th} resolution, $\mathbf{R}_i \in \mathbb{R}^3$ is a resolution-specific scaling factor, and $T \in \mathbb{N}$ is the size of the hash table. Each hashed index \mathbf{h}_i points to a feature vector in the hash table, which is then used to interpolate the final feature representation for the input point \mathbf{p} . The summary of hash-grid-based NeRFs is provided in Table XI. In the following, we review these models for camera, Lidar, and joint camera and Lidar simulation.

1) *RGB Image Synthesis*: The pioneering work, MINE [113], introduces a continuous depth generalisation of the MultiPlane Images (MPI) approach by integrating NeRFs. This method performs dense 3D reconstruction and novel view synthesis from a single image using a hash-grid encoding. By predicting RGB and volume density values at arbitrary depth planes, MINE effectively reconstructs the 3D scene within the camera frustum, filling in occluded contents.

SUDS [114] extends NeRFs to dynamic large-scale urban scenes, addressing the limitations of prior methods that typically reconstruct single video clips of short durations. SUDS introduces a scalable representation by factorising scenes into three separate hash table data structures to efficiently encode static, dynamic, and far-field radiance fields. This approach uses unlabelled target signals, including RGB images, sparse Lidar, self-supervised 2D descriptors, and optical flow, enabling photometric, geometric, and feature-metric reconstruction losses. SUDS decomposes dynamic scenes into static backgrounds, individual objects, and their motions, scaling up to tens of thousands of objects across 1.2 million frames from 1700 videos.

StreetSurf [116] presents a multi-view implicit surface reconstruction technique tailored for street view images. Using hash-grid encoding and geometric priors from monocular models, this method addresses the challenges posed by unbounded street views captured with non-object-centric, long, and narrow camera trajectories. The key contribution includes dividing the scene into close-range, distant-view, and sky regions with aligned cuboid boundaries, facilitating fine-grained and disentangled representation.

MARS [117] introduces an instance-aware, modular, and realistic simulator for ADS, leveraging hash-grid encoding to model foreground instances and background environments separately, as shown in Fig. 12. This modular framework supports flexible switching between different NeRF-related backbones, sampling strategies, and input modalities. MARS achieves state-of-the-art photo-realism by decomposing scenes into foreground and background components, allowing for independent control over static and dynamic properties of instances.

EmerNeRF [118] introduces a self-supervised approach for learning spatial-temporal representations of dynamic driving scenes. This method stratifies scenes into static and dynamic

TABLE XI: Summary of hash-grid-based NeRFs

Model	Sensor	Year	Description	Datasets
MINE [113]	Camera	2021	Performs dense 3D reconstruction and novel view synthesis from a single image using a hash-grid encoding.	KITTI [19]
SUDS [114]	Camera	2023	Extends NeRFs to dynamic large-scale urban scenes, factorising scenes into separate hash table data structures for efficient encoding.	KITTI [19], V-KITTI 2 [115]
StreetSurf [116]	Camera	2023	Multi-view implicit surface reconstruction technique tailored for street view images, dividing the scene into close-range, distant-view, and sky regions.	Waymo [33]
MARS [117]	Camera	2023	Instance-aware, modular, and realistic simulator for autonomous driving, leveraging hash-grid encoding to model foreground instances and background environments separately.	KITTI [19], V-KITTI 2 [115]
EmerNeRF [118]	Camera	2023	Self-supervised approach for learning spatial-temporal representations of dynamic driving scenes, capturing scene geometry, appearance, motion, and semantics.	NOTR [118]
DGNR [119]	Camera	2023	Density-Guided Neural Rendering learns a density space to guide the construction of a point-based renderer, eliminating the need for geometric priors.	KITTI [19]
LightSim [120]	Camera	2023	Neural lighting simulation system for urban driving scenes, generating diverse, controllable, and realistic camera data with accurate illumination and shadows.	PandaSet [99]
NeRF-LiDAR [121]	Lidar	2023	Leverages real-world information to generate realistic Lidar point clouds by reconstructing 3D scenes using multi-view images and sparse Lidar data.	nuScenes [55]
LiDAR-NeRF [122]	Lidar	2023	Differentiable end-to-end Lidar rendering framework that synthesises novel views for Lidar sensors, learning geometry and attributes of 3D points.	KITTI-360 [69], NeRF-MVL [122]
NFL [123]	Lidar	2023	Optimises a neural field scene representation from Lidar measurements to synthesise realistic Lidar scans from novel viewpoints.	Waymo [33]
DyNFL [124]	Lidar	2023	Neural field-based approach for high-fidelity re-simulation of Lidar scans in dynamic driving scenes, constructing an editable neural field.	Waymo [33], KITTI [19]
UniSim [125]	Camera & Lidar	2023	Closed-loop neural sensor simulation system for self-driving, utilising hash-grid encoding to build neural feature grids for reconstructing scenes.	PandaSet [99]
NeuRAD [126]	Camera & Lidar	2023	Neural rendering method designed for dynamic automotive scenes, with extensive sensor modelling for both cameras and Lidar.	nuScenes [55], PandaSet [99], Argoverse 2 [105], KITTI [19], ZOD [127]
AlignMiF [128]	Camera & Lidar	2024	Geometry-aligned multimodal implicit field for joint Lidar-camera synthesis, implementing Geometry-Aware Alignment (GAA) and Shared Geometry Initialisation (SGI).	KITTI-360 [69], Waymo [33]

fields and further parameterises an induced flow field from the dynamic field to aggregate multi-frame features, enhancing the rendering precision of dynamic objects. EmerNeRF’s key contribution is its ability to capture scene geometry, appearance, motion, and semantics without relying on ground-truth annotations or pre-trained models.

DGNR [119] introduces Density-Guided Neural Rendering, which learns a density space from scenes to guide the construction of a point-based renderer. This method eliminates the need for geometric priors by intrinsically learning them from the density space through volumetric rendering. DGNR employs a differentiable renderer to synthesise images from neural density features and uses a density-based fusion module along with geometric regularisation to optimise the density space.

LightSim [120] presents a neural lighting simulation system for urban driving scenes, enabling the generation of diverse, controllable, and realistic camera data. LightSim reconstructs lighting-aware digital twins from real-world sensor data, including geometry, appearance, and estimated scene lighting. This system facilitates actor insertion, modification, removal, and relighting from new viewpoints with accurate illumination and shadows.

2) Lidar Point Cloud Synthesis: NeRF-LiDAR [121] presents a novel Lidar simulation method that leverages real-world information to generate realistic Lidar point clouds. This method reconstructs 3D scenes using multi-view images and sparse Lidar data collected by self-driving cars. By learning the NeRF representation for real-world scenes, NeRF-LiDAR generates point clouds with accurate semantic labels, which significantly boosts the performance of 3D segmentation models trained on this simulated data.

LiDAR-NeRF [122] introduces the first differentiable end-to-end Lidar rendering framework, designed to synthesise novel views for Lidar sensors. This method leverages neural radiance fields to jointly learn geometry and attributes of 3D points, such as intensity and ray-drop probability, without explicit 3D reconstruction. The approach incorporates structural regularisation to preserve local details, significantly improving the accuracy of synthesised Lidar patterns.

NFL model [123] optimises a neural field scene representation from Lidar measurements to synthesise realistic Lidar scans from novel viewpoints. NFL combines the rendering power of neural fields with a detailed, physically motivated model of the Lidar sensing process, accurately reproducing key sensor behaviours like beam divergence, secondary returns,

TABLE XII: Summary of point-based and multi-planar NeRFs

Model	Sensor	Year	Description	Datasets
NPLF [130]	Camera	2021	Encodes light fields on sparse point clouds, enabling efficient novel view synthesis for large-scale driving scenarios.	Waymo [33]
Chang et al. [131]	Camera	2023	Integrates Lidar maps and 2D conditional GANs to improve novel view synthesis in outdoor environments.	Argoverse 2 [105]
DNMP [132]	Camera	2023	Mesh-based rendering combined with neural representations for efficient urban-level radiance field construction.	KITTI-360 [69], Waymo [33]
LiDAR4D [124]	Lidar	2024	Differentiable Lidar-only framework for novel space-time Lidar view synthesis, using a 4D hybrid representation.	KITTI-360 [69], NuScenes [55]

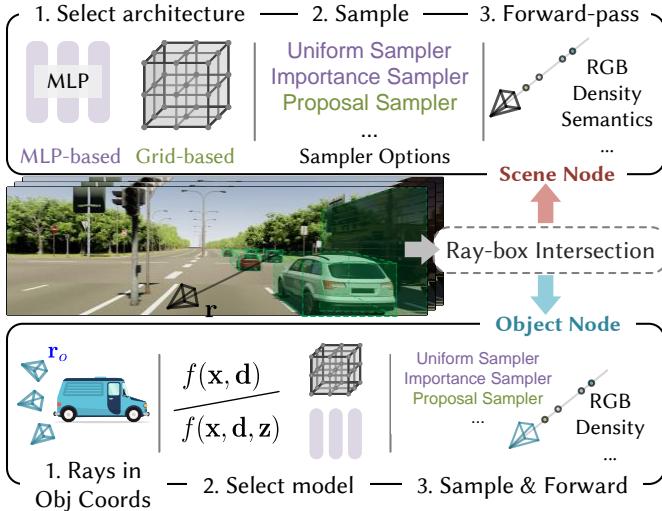


Fig. 12: Decomposed ray intersection pipeline of MARS [117], a hash-grid-based NeRF for RGB image synthesis.

and ray dropping. The key innovation lies in integrating these physical characteristics into the neural field framework, leading to improved realism in synthesising Lidar views.

DyNFL model [124] introduces a novel neural field-based approach for high-fidelity re-simulation of Lidar scans in dynamic driving scenes. This method processes Lidar measurements from dynamic environments, constructing an editable neural field by separately reconstructing static backgrounds and dynamic objects. DyNFL utilises a neural field composition technique that integrates reconstructed neural assets from various scenes through a ray drop test, accounting for occlusions and transparent surfaces. This approach significantly improves the physical fidelity and flexible editing capabilities of dynamic scene Lidar simulation.

3) *Joint Camera and Lidar Simulation*: UniSim [125] presents a closed-loop neural sensor simulation system for self-driving, utilising hash-grid encoding to build neural feature grids that reconstruct both static backgrounds and dynamic actors in the scene. This method enables the generation of realistic multi-sensor simulations from a single recorded log, allowing for scene manipulation and the creation of new scenarios. UniSim integrates learnable priors for dynamic objects and a convolutional network to handle unseen regions, improving the realism of extrapolated views.

NeuRAD [126] is a neural rendering method specifically

designed for dynamic automotive scenes. It features a simple network design and extensive sensor modelling for both cameras and Lidar, including effects such as rolling shutter, beam divergence, and ray dropping. NeuRAD employs a unified neural feature field for both static and dynamic elements, enabling realistic sensor data generation and the editing of the pose of the ego vehicle and other road users.

AlignMiF [128] introduces a geometry-aligned multimodal implicit field for joint Lidar-camera synthesis, addressing the challenge of misalignment between different sensor modalities. By implementing Geometry-Aware Alignment (GAA) and Shared Geometry Initialisation (SGI), AlignMiF aligns the coarse geometry across Lidar and camera data, significantly enhancing the fusion process.

C. Point-based and Multi-planar Encoding

Point-based and multi-planar NeRFs are advanced techniques designed to improve the efficiency and quality of scene representation and reconstruction in NeRFs. In point-based NeRFs, the scene is represented using a sparse set of 3D points, often captured by Lidar sensors. Each point is associated with feature descriptors that are used to interpolate light fields for rendering. This approach reduces the computational burden associated with traditional volumetric models and enables efficient processing of large-scale environments. Multi-planar encoding extends this concept by utilising multiple planes to represent the scene at different depths. The summary of point-based and multi-planar NeRFs is provided in Table XII. In the following, we review these models for camera and Lidar simulation.

NPLF [130] introduces a novel representation that encodes light fields on sparse point clouds, enabling efficient novel view synthesis for large-scale driving scenarios. This method departs from traditional volumetric models by leveraging the geometric properties of point clouds, which are sparsely captured by Lidar sensors. NPLF computes per-point features using a learned feature extractor and interpolates these features to form light field descriptors for each ray, ensuring consistent and unique descriptions. This technique significantly reduces the computational burden of volumetric sampling and achieves realistic novel view synthesis with minimal training data.

Chang et al. [131] introduce a point-based NeRF framework that integrates Lidar maps and 2D conditional GANs (cGANs) to improve novel view synthesis in outdoor environments. This method leverages the strong 3D geometry priors provided by Lidar sensors, significantly enhancing ray sampling locality

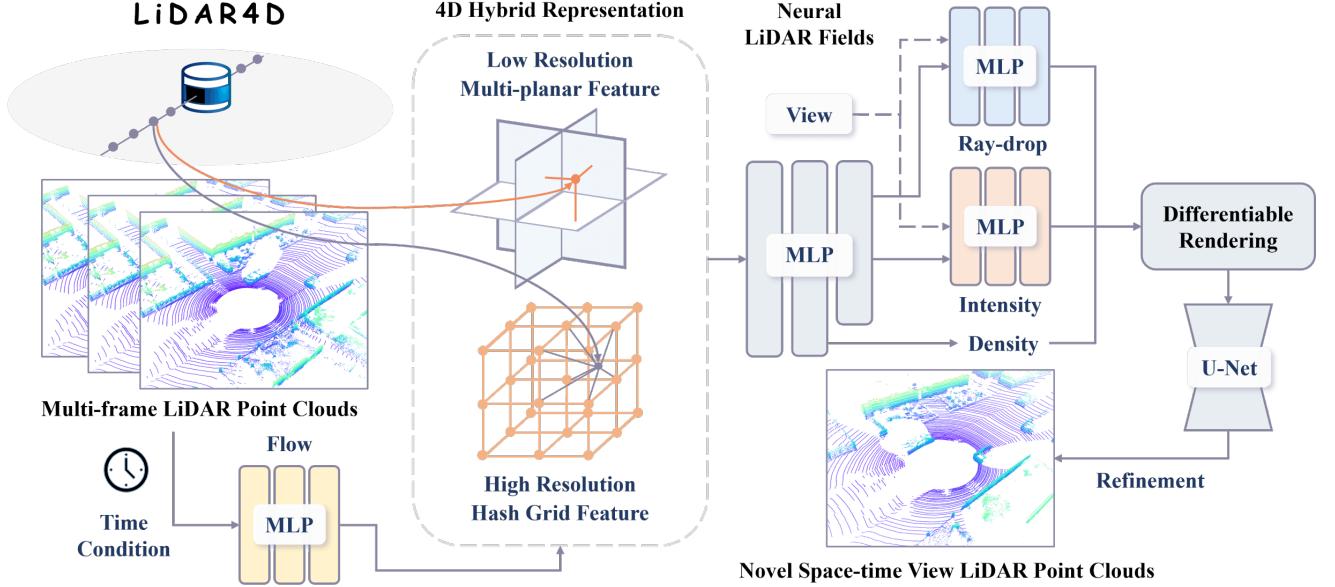


Fig. 13: Overview of LiDAR4D [124] model, a NeRF model which uses both multi-planar and hash-grid features for Lidar point cloud simulation.

TABLE XIII: Summary of 3D Gaussian splatting models (all used for camera RGB image synthesis).

Model	Year	Description	Datasets
PVG [133]	2023	Unified representation model capturing static and dynamic elements using periodic vibration-based temporal dynamics and adaptive control for efficient large scene representation.	Waymo [33], KITTI [19]
Street Gaussians [134]	2023	Explicit scene representation modelling dynamic urban streets from monocular videos, separating foreground vehicles from the static background, enabling real-time rendering.	KITTI, Waymo [33]
DrivingGaussian [135]	2024	Efficient framework for reconstructing and rendering dynamic scenes, using incremental static 3D Gaussians and a composite dynamic Gaussian graph for multiple moving objects.	nuScenes [55], KITTI-360 [69]
HUGS [136]	2024	Comprehensive framework for urban scene understanding via 3D Gaussian splatting, optimising geometry, appearance, semantics, and motion for real-time novel view synthesis.	KITTI [19], KITTI-360 [69], V-KITTI 2 [115]
SGD [137]	2024	Enhances street view synthesis by leveraging a diffusion model for prior knowledge and 3D Gaussian splatting to improve rendering quality from sparse training views.	KITTI [19], KITTI-360 [69]
DeSiRe-GS [138]	2024	Self-supervised 4D Gaussian splatting model decomposing static and dynamic elements with temporal consistency and motion masks for accurate scene reconstruction.	KITTI [19], Waymo [33]

and reducing artefacts caused by imperfect Lidar data. By using Lidar maps as sparse samples of the environment, the system assigns localised embeddings and performs volume rendering. The integration of cGANs refines the synthesised images, producing higher quality outputs.

Lu et al. [132] propose a new framework, DNMPs, for efficient urban-level radiance field construction. DNMPs leverage mesh-based rendering combined with neural representations to model both geometry and radiance information compactly. Each DNMP consists of connected deformable mesh vertices paired with radiance features, optimising shape and radiance using a latent code. This approach significantly reduces computational costs and memory usage while maintaining high-quality rendering.

LiDAR4D [124] introduces a differentiable Lidar-only framework for novel space-time Lidar view synthesis. This method tackles the challenges of dynamic scene reconstruction by using a 4D hybrid representation that combines multi-planar and grid features, as shown in Fig. 13. LiDAR4D

incorporates geometric constraints derived from point clouds to improve temporal consistency and uses global optimisation of ray-drop probability to preserve cross-region patterns.

VII. VOLUME RENDERERS – 3D GAUSSIAN SPLATTING

3D Gaussian Splatting (3DGS) models have emerged as a powerful technique for scene rendering in ADS. These models utilise 3D Gaussian functions to efficiently capture and represent both static and dynamic elements in urban environments. By incorporating mechanisms that enhance temporal continuity and leverage adaptive control strategies, 3DGS models ensure detailed and consistent scene representations. They employ point-based encoding of the scene, using incremental and composite approaches to model static backgrounds and dynamic objects. Compared to NeRFs, 3DGS offers significant advantages in rendering speed, achieving up to 900-fold acceleration [133], which is crucial for real-time ADS applications. It also demonstrates superior performance in reconstruction quality and novel view synthesis, particularly

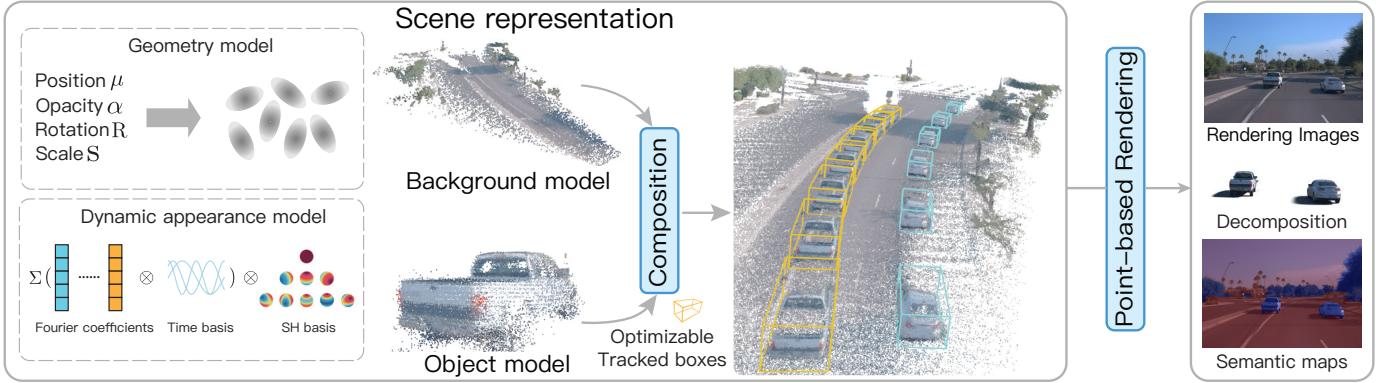


Fig. 14: Overview of rendering pipeline in Street Gaussians [134], a volume renderer based on 3D Gaussian Splattting for RGB image synthesis.

in dynamic urban scenes, as evidenced by improved PSNR, SSIM, and LPIPS metrics. However, 3DGS faces challenges in maintaining temporal consistency in dynamic scenes and may struggle with fine geometric details, areas where NeRFs might perform better due to their continuous representation. These trade-offs between rendering speed, image quality, and geometric accuracy are important considerations when applying 3DGS models in ADS scenarios. The summary of 3DGS models is provided in Table XIII. In the following, we review these models that are all used to synthesise RGB images.

PVG method [133] introduces a unified representation model designed to efficiently and uniformly capture both static and dynamic elements in large-scale urban scenes. Building on the 3D Gaussian splatting technique, PVG incorporates periodic vibration-based temporal dynamics, enabling a cohesive representation of scene characteristics. The model introduces a temporal smoothing mechanism and a position-aware adaptive control strategy to enhance temporal continuity and large scene representation learning with sparse training data.

DrivingGaussian [135] presents an efficient framework for reconstructing and rendering surrounding dynamic scenes in autonomous driving contexts. This method sequentially models static backgrounds with incremental static 3D Gaussians and uses a composite dynamic Gaussian graph to handle multiple moving objects. By leveraging Lidar priors for Gaussian splatting, DrivingGaussian reconstructs scenes with high detail and maintains panoramic consistency across multi-camera setups. The model outperforms existing methods in dynamic driving scene reconstruction, achieving photo-realistic surround-view synthesis with high fidelity and consistency.

Street Gaussians [134] introduces an explicit scene representation designed to efficiently model dynamic urban street scenes from monocular videos. This method represents dynamic urban streets as point clouds equipped with semantic logits and 3D Gaussians, separating foreground vehicles from the static background (see Fig. 14). By optimising tracked poses and utilising a dynamic spherical harmonics model for the dynamic appearance of vehicles, Street Gaussians can efficiently compose object vehicles and backgrounds.

HUGS [136] proposes a comprehensive framework for urban scene understanding from RGB images. By utilising

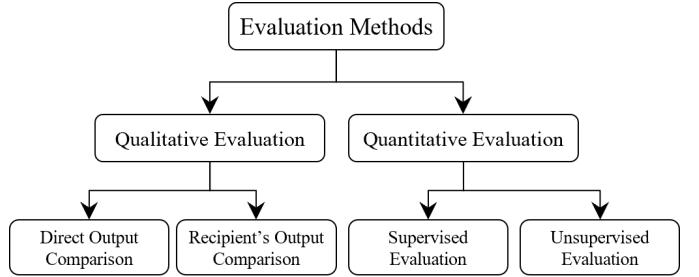


Fig. 15: Categorising validation approaches of perception sensor simulation models.

3DGS, HUGS jointly optimises geometry, appearance, semantics, and motion, facilitating real-time novel view synthesis and dynamic scene reconstruction. This method incorporates physical constraints for moving object poses, improving accuracy even with noisy 3D bounding box predictions.

Yu et al. [137] present a novel approach for enhancing street view synthesis in ADS simulations. This method leverages a DM to provide prior knowledge, combined with 3DGS to improve rendering quality from sparse training views. By fine-tuning the DM with images from adjacent frames and depth data from Lidar point clouds, the model regularises 3DGS training and enhances the quality of novel view synthesis.

DeSiRe-GS [138] extends 3D Gaussian splatting to a 4D representation for self-supervised static-dynamic decomposition and surface reconstruction in urban driving scenarios. By leveraging 2D motion masks and temporal consistency, it models dynamic elements as time-varying variables and reconstructs Gaussian surfaces without explicit 3D annotations. The method achieves state-of-the-art scene decomposition and reconstruction performance, effectively addressing data sparsity challenges.

VIII. EVALUATION APPROACHES

Similar to any other modelling approach, it is crucial to thoroughly evaluate the performance of simulation models concerning their intended role within the system [139]. In the evaluation of perception sensor simulation models for ADS, two general methodologies, namely qualitative and

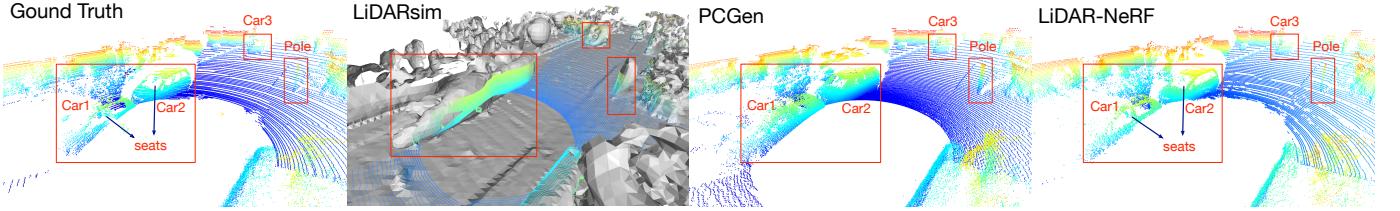


Fig. 16: Qualitative evaluation of a Lidar simulation model by direct comparison. The synthesised point cloud of LiDAR-NeRF [122] is compared to state-of-the-art models and ground-truth.

quantitative evaluation, are commonly employed. Qualitative evaluation involves visualising either the output of the model itself or the output of a recipient component, such as a perception model. On the other hand, quantitative evaluation entails assessing the model based on either supervised metrics, when reference data is available, or unsupervised metrics, when reference data is not accessible. The categorisation of the evaluation approaches is depicted in Fig. 15. In the subsections below, we will explore existing examples of both qualitative and quantitative evaluations in the literature of camera and Lidar simulation models for ADS.

A. Qualitative Evaluation

Qualitative evaluation is the most common approach for assessing the sensor simulation models. This process involves visualising the output of the sensor model and comparing it to state-of-the-art methods and real data. Moreover, some studies emphasise visualising the output of a perception model, such as a semantic segmentation network. A more detailed exploration of these two qualitative evaluation approaches is provided in the subsequent paragraphs.

1) Direct Output Visualisation: Most camera and Lidar models visualise their synthesised outputs and compare them with real data and state-of-the-art methods. The input data for sensor simulation models is also visualised (except in unconditional approaches) to understand the information the model learns from. For example, in the sim-to-real mapping network introduced by [56], various simulated buffers such as normal, depth, and albedo, are processed and then visualised during the evaluation phase. For comparison with real sensory data, the model’s output is either compared to ground-truth data in cases of supervised methods (see Fig. 16), or with nearest neighbour samples from the training dataset regarding unsupervised approaches. In examples such as the LidarSim [93] framework, the entire scene is reconstructed using real data, providing corresponding ground-truth for each Lidar viewpoint. Conversely, Dusty [24] generates data randomly, and the nearest neighbour samples from real data are used for comparison. Regarding data representation, camera models often visualise synthesised RGB images, whereas Lidar models visualise BEV images, range images or 3D point clouds based on the utilised data representation.

Additionally, learned models are sometimes applied to applications such as data restoration, and the results are visualised to demonstrate the model’s effectiveness. For example, Dusty2 [25] model was used to reconstruct real range

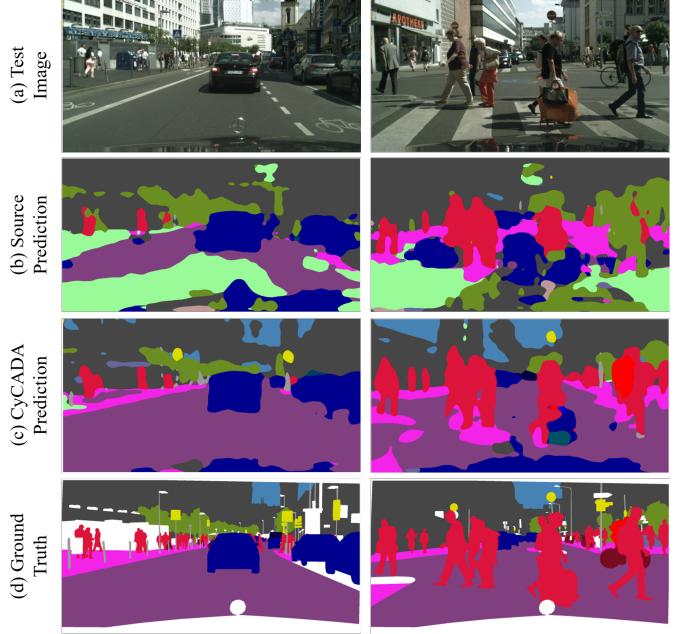


Fig. 17: Qualitative validation of CyCADA [40] by comparing the output of a downstream task, e.g. semantic segmentation network. The semantic segmentation network is trained with two different image sources: (1) synthetic images of the GTA-V dataset, and (2) the modelled ones by CyCADA. The test images (a) are fed to these networks and the prediction is depicted in (b) and (c), respectively. The last row (d) shows the ground-truth layout.

images or restore distorted images. The high-quality restored or reconstructed images demonstrate the model’s ability to cover data distribution effectively.

2) Recipient’s Output Comparison: As perception sensors are not standalone systems in ADS, it is essential to consider their integration with downstream components that receive sensory data. Considering a data-driven model as the recipient component, several studies [40], [92], [95], [93] have visualised the predicted annotations of the models. This approach allows for evaluating the performance of the models that have been trained on synthesised data using real test data. The closer the output of the sensor model resembles the real-world data, the higher the accuracy of the perception model in detecting and interpreting its input. Fig. 17 demonstrates an example of CyCADA’s [40] evaluation through predictions of a semantic segmentation network, FCN [140]. As shown, the FCN model

is much more accurate when trained on CyCADA’s synthesised images than on the source images. This demonstrates the effectiveness of CyCADA in improving the realism of GTA-V simulated images.

B. Quantitative Evaluation

Quantitative evaluation of the sensor simulation models entails establishing a mathematical metric to measure the model’s performance against the expected output. We categorise these quantitative evaluation methods into supervised and unsupervised groups. Supervised evaluation techniques require paired model input and ground-truth output for the assessment, while unsupervised evaluation techniques do not enforce such necessity. We will explain the quantitative evaluation techniques in more detail in the following paragraphs.

1) *Supervised Evaluation*: To quantitatively evaluate a sensor model in a supervised manner, it is necessary to establish correspondence between the synthesised sensory data and the real-world data. The supervised evaluation methods are often utilised in supervised modelling approaches where the models are trained with a pair of input-output data, thus enabling comparison with ground-truth. For instance, in the cases where semantic summation layout is transformed into RGB images [27] or Lidar intensity is predicted using the spatial coordinates [92], the ground-truth RGB or intensity image in the test set is used for the supervised evaluation. Similarly, other frameworks such as SurfelGAN [32] and LidarSim [93] reconstruct the 3D scene using real sensory data frames, thus providing access to the expected models’ output at specific viewpoints of the sensor.

Supervised evaluation metrics commonly focus on calculating the average per-pixel error in image-based sensory data representations. These metrics include Root Mean Squared Error (RMSE, i.e. L2 distance error), Mean Absolute Error (MAE, L1 distance error), thresholded accuracy (ratio of pixels with the error less than a certain threshold), and PSNR. For calculating the distance of the Lidar point cloud in 3D representation, Chamfer Distance (CD) and Earth Mover’s Distance (EMD) [141] are frequently utilised. Additionally, other supervised metrics such as SSIM and LPIPS [142] are used, relying on structural and feature similarity rather than solely on pixel-level error.

2) *Unsupervised Evaluation*: Building an environment model that accurately reflects complex real-world scenarios is a challenging task. Therefore, there is often no direct correspondence between real and synthesised sensory data, thus unsupervised evaluation methods are commonly employed. These methods typically rely on constructing two sets of synthesised and real data, considering them as distributions, and employing metrics to calculate the distance between these distributions. Frechet Inception Distance (FID) [143], Kernel Inception Distance (KID) [144], and Sliced Wasserstein Distance (SWD) [145] are common metrics used to measure this distance for the sensory data represented as images. In the work by Ritcher et al. [56], the authors proposed semantically aligned Kernel VGG distance (sKVD) that addresses the bias toward semantic similarity of the scenes, a limitation often

TABLE XIV: Summary of quantitative evaluation metrics.

Category	Description	Metric
	pixel-level error measurement	RMSE, MAE, PSNR
Supervised	perceptual similarity	SSIM, LPIPS [142]
	3D point cloud distance	CD, EMD [141]
	distribution distance (images)	FID [143], KID [144], sKVD [56], SWD [145]
Unsupervised	distribution distance (point clouds)	JSD [141], MMD [141]
	distribution diversity	COV [141], 1-NNA [141]
	perception model’s performance	IOU, mAP, pixAcc, classAcc

found in FID/KID. Concerning finding the distance between distributions of two Lidar 3D point clouds, Jensen-Shannon Divergence (JSD) and Minimum Matching distance (MMD) have also been employed [24]. In assessing the performance of unconditional modelling approaches, metrics such as Coverage (COV) and 1-Nearest Neighbour Accuracy (1-NNA) have also been used [24] to measure the diversity of generated samples.

Some studies have also incorporated human perceptual studies [91] to measure the realism of the synthesised RGB images. This evaluation involves visualising the paired synthesised and real images to the human users and reporting the percentage of times when users preferred the synthesised image. The studies commonly used the Amazon Mechanical Turk (AMT) platform for conducting evaluations.

The most widely used unsupervised evaluation approach found in the literature involves assessing the performance of a downstream perception model. This process starts by conducting a set of synthesised sensory data to train the perception model. After training with the synthesised dataset, the perception model is then validated on a real sensory dataset. The performance of the perception model is compared to its performance when initially trained on real data. The principle of this approach is that the closer the perception model’s performance to its performance on real data, the more realistic the synthesised data is perceived by the model. Typically, the perception models used in these evaluations are SOTA object detection or semantic segmentation models. Key performance metrics for these models include Intersection Over Union (IOU), average Precision (AP), average pixel accuracy (pixAcc), and average class accuracy (classAcc). The summary of the quantitative evaluation metrics is provided in Table XIV.

IX. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This review explored state-of-the-art data-driven camera and Lidar simulation models for ADS development and validation. The paper categorised and reviewed 96 different models from the novel perspective of generative models, including GANs, diffusion models, and auto-regressive models, as well as volume renderers comprising NeRFs and 3DGs. By focusing on these data-driven approaches, the review aimed to provide insights into the rapid expansion of the literature in this field, driven by the growing availability of real-world recorded perception sensory datasets and the success of deep learning

models in synthesising high-dimensional data. While these models have shown promising results in enhancing the realism and efficiency of sensor simulation, several challenges and opportunities for future research have emerged.

Standardising Evaluation and Benchmarking: A critical challenge in current research is the lack of standardised protocols and comprehensive benchmarks for assessing sensor simulation models in ADS applications. To address this limitation, future research should prioritise:

- Developing robust benchmarks to evaluate simulation models' capability in generating data for edge cases, such as adverse weather conditions, crucial for ADS safety testing.
- Developing standardised protocols to evaluate the computational efficiency and scalability of simulation models, focusing on their suitability for real-time ADS applications and large-scale data generation.
- Enhancing evaluation methodologies beyond traditional ego-vehicle trajectory analysis. This expansion should include assessments of novel spatial and temporal view synthesis capabilities, particularly for volume renderers [124].

Enhancing Real-time Performance and Controllability:

The suitability of data-driven models for real-time ADS applications, such as Hardware-in-the-Loop (HiL) testing, remains largely unexplored in the literature. This gap is particularly evident in certain diffusion models [73], which currently fall short of real-time capabilities. Furthermore, most of the discussed generative models are either unconditional or rely on simplistic control inputs, which restricts their adequacy for real-world applications, e.g. implementing diverse testing scenarios. These limitations present significant challenges for integrating data-driven models into time-critical ADS testing workflows. To address these issues, future research should focus on:

- Optimising deep learning-based models to achieve real-time or faster-than-real-time performance, potentially through hardware-specific implementations and algorithmic innovations.
- Accelerating generation algorithms in computationally intensive models, including diffusion models, leveraging techniques such as latent consistency [146].
- Enhancing the controllability of generative models to allow for precise adjustments to synthesised sensory data, enabling researchers to replicate specific testing scenarios, such as varying lighting, weather conditions, and sensor occlusions.

Improving Trustworthiness and Generalisation: Deep learning models inherently struggle to generalise beyond their training data distribution, raising concerns about their reliability in diverse scenarios [147]. This limitation is particularly pertinent to reviewed data-driven simulation models, which heavily rely on deep learning approaches. Given that these models play a crucial role in ADS testing, their ability to accurately represent a wide range of driving conditions is of paramount importance. To enhance trustworthiness, future research should prioritise:

- Exploring hybrid approaches that combine data-driven simulation methods with physics-based models to improve the generalisation capabilities and overall reliability.
- Investigating transfer learning approaches to leverage knowledge from high-fidelity and complex models to improve the performance of more efficient, deployable models.
- Developing robust methods to quantify and communicate uncertainty in simulated sensor data, enhancing the transparency and reliability of virtual ADS testing environments.

Enhancing Volume Rendering Techniques: Despite significant advancements, volume renderers [126], [136], [133], [125], [124] still face several specific challenges in accurately simulating complex driving scenes. To address these limitations, future research should focus on:

- Enhancing NeRFs to accurately simulate deformable actors and dynamic objects over extended periods.
- Improving 3DGS performance in challenging environmental conditions and night-time scenarios.
- Developing techniques for modelling dynamic scene elements and enhancing temporal coherence through scene flow estimation [133].
- Enhancing geometric accuracy for both static and dynamic objects, particularly in handling occlusions and long-distance motions.

Data-driven camera and Lidar simulation models have made significant strides, yet critical areas for improvement remain. Addressing these challenges is crucial for enhancing the reliability, efficiency, and fidelity of sensor simulation models in ADS applications. As the field evolves, interdisciplinary collaboration and the development of standardised evaluation methods will be key to unlocking the full potential of these technologies and ensuring their safe, effective deployment in real-world ADS scenarios.

ACKNOWLEDGMENT

This research is supported in part by the University of Warwick's Centre for Doctoral Training in Future Mobility Technologies and in part by the Hi-Drive Project through the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No 101006664. We thank Dr Christoph Kessler from Ford for his valuable comments and suggestions on improving the paper. The sole responsibility of this publication lies with the authors.

REFERENCES

- [1] Uber, "Uber in fatal crash had safety flaws say us investigators," Available at <https://www.bbc.co.uk/news/business-50312340> (accessed: 14.07.2020).
- [2] Tesla, "Tesla in fatal california crash was on autopilot," Available at <https://www.bbc.co.uk/news/world-us-canada-43604440> (accessed: 14.07.2020).
- [3] N. Kalra and S. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 12 2016.

- [4] S. M. Grigorescu, B. Trasnea, T. T. Cocias, and G. Macesanu, “A survey of deep learning techniques for autonomous driving,” *Journal of Field Robotics*, vol. 37, pp. 362 – 386, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:204744017>
- [5] W. Liu, Q. Dong, P. Wang, G. Yang, L. Meng, Y. Song, Y. Shi, and Y. Xue, “A survey on autonomous driving datasets,” in *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, 2021, pp. 399–407.
- [6] Schläger, Muckenhuber, Schmid-Schläger, and Holzer, “State-of-the-art sensor models for virtual testing of advanced driver assistance systems/autonomous driving functions,” *SAE Intl J*, 2020.
- [7] Z. Liu, S. Minghao, J. Zhang, S. Liu, H. Blasinski, T. Lian, and B. Wandell, “A system for generating complex physically accurate sensor images for automotive applications,” *Electronic Imaging*, vol. 2019, pp. 53–1, 01 2019.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf
- [9] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.
- [10] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: representing scenes as neural radiance fields for view synthesis,” *Commun. ACM*, vol. 65, no. 1, p. 99–106, dec 2021. [Online]. Available: <https://doi.org/10.1145/3503250>
- [11] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023. [Online]. Available: <https://doi.org/10.1145/3592433>
- [12] F. Rosique, P. Navarro Lorente, C. Fernandez, and A. Padilla, “A systematic review of perception system and simulators for autonomous vehicles research,” *Sensors*, vol. 19, p. 648, 02 2019.
- [13] Y. Kang, H. Yin, and C. Berger, “Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments,” *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 171–185, 2019.
- [14] A. Elmquist and D. Negruț, “Modeling cameras for autonomous vehicle and robot simulation: An overview,” *IEEE Sensors Journal*, vol. 21, no. 22, pp. 25 547–25 560, 2021.
- [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.
- [18] L. Caccia, H. van Hoof, A. C. Courville, and J. Pineau, “Deep generative modeling of lidar data,” in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5034–5040, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:54445260>
- [19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *International Journal of Robotics Research (IJRR)*, 2013.
- [20] S. Azadi, M. T. Tschannen, E. Tzeng, S. Gelly, T. Darrell, and M. Lucić, “Semantic bottleneck scene generation,” *Tech. Rep.*, 2019.
- [21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [22] A. Volokitin, E. Konukoglu, and L. Van Gool, “Decomposing image generation into layout prediction and conditional synthesis,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1530–1538.
- [23] G. L. Moing, T. Vu, H. Jain, P. Perez, and M. Cord, “Semantic palette: Guiding scene generation with class proportions,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2021, pp. 9338–9346. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00922>
- [24] K. Nakashima and R. Kurazume, “Learning to drop points for lidar scan synthesis,” in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 222–229, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:232035774>
- [25] K. Nakashima, Y. Iwashita, and R. Kurazume, “Generative range imaging for learning scene priors of 3d lidar data,” 2022.
- [26] G. Eskandar, Y. Farag, T. Yenamandra, D. Cremers, K. Guirguis, and B. Yang, “Urban-stylegan: Learning to generate and manipulate images of urban scenes,” 2023.
- [27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:6200260>
- [28] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:41805341>
- [29] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2332–2341.
- [30] X. Liu, G. Yin, J. Shao, X. Wang, and H. Li, *Learning to predict layout-to-image conditional convolutions for semantic image synthesis*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [31] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, “Sean: Image synthesis with semantic region-adaptive normalization,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [32] Z. Yang, Y. Chai, D. Anguelov, Y. Zhou, P. Sun, D. Erhan, S. Rafferty, and H. Kretzschmar, “Surfegan: Synthesizing realistic sensor data for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [33] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [34] E. Schönfeld, V. Sushko, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, “You only need adversarial supervision for semantic image synthesis,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=yvQKLaqNE6M>
- [35] H. Tang, X. Qi, G. Sun, D. Xu, N. Sebe, R. Timofte, and L. Van Gool, “Edge guided gans with contrastive learning for semantic image synthesis,” *ICLR*, 2023.
- [36] M. Hariati, O. Laurent, R. Kazmierczak, S. Zhang, A. Bursuc, A. Yao, and G. Franchi, “Learning to generate training datasets for robust semantic segmentation,” 2023.
- [37] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.
- [38] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” 2020.
- [39] S. Benaim and L. Wolf, “One-sided unsupervised domain mapping,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 752–762.
- [40] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, “CyCADA: Cycle-consistent adversarial domain adaptation,” in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1989–1998.
- [41] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, “Multimodal unsupervised image-to-image translation,” in *ECCV*, 2018.
- [42] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, “Contrastive learning for unpaired image-to-image translation,” in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Berlin, Heidelberg: Springer-Verlag, 2020, p. 319–345. [Online]. Available: https://doi.org/10.1007/978-3-030-58545-7_19
- [43] Z. Jia, B. Yuan, K. Wang, H. Wu, D. Clifford, Z. Yuan, and H. Su, “Semi-robust unpaired image translation for data with unmatched

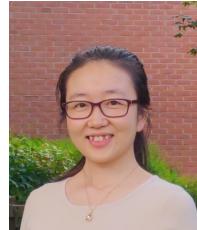
- semantics statistics,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 273–14 283.
- [44] C. Zheng, T.-J. Cham, and J. Cai, “The spatially-correlative loss for various image translation tasks,” 06 2021, pp. 16 402–16 412.
- [45] C. Jung, G. Kwon, and J. C. Ye, “Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 260–18 269.
- [46] S. Xie, Y. Xu, M. Gong, and K. Zhang, “Unpaired image-to-image translation with shortest path regularization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 177–10 187.
- [47] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in *European Conference on Computer Vision (ECCV)*, ser. LNCS, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer International Publishing, 2016, pp. 102–118.
- [48] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [49] K. Saleh, A. Abobakr, M. Attia, J. Iskander, D. Nahavandi, M. Hossny, and S. Nahvandi, “Domain adaptation for vehicle detection from bird’s eye view lidar point cloud data,” in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3235–3242.
- [50] A. Dosovitskiy, G. Ros, F. Codevilla, A. M. López, and V. Koltun, “Carla: An open urban driving simulator,” *ArXiv*, vol. abs/1711.03938, 2017.
- [51] A. Carlson, K. A. Skinner, R. Vasudevan, and M. Johnson-Roberson, “Sensor transfer: Learning optimal sensor effect image augmentation for sim-to-real domain adaptation,” *IEEE Robotics and Automation Letters*, vol. 4, pp. 2431–2438, 2018.
- [52] R. Gong, D. Dai, Y. Chen, W. Li, and L. V. Gool, “Analogical image translation for fog generation,” *arXiv*, 6 2020. [Online]. Available: <http://arxiv.org/abs/2006.15618>
- [53] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, “Virtual worlds as proxy for multi-object tracking analysis,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4340–4349.
- [54] M. Tremblay, S. S. Halder, R. de Charette, and J. F. Lalonde, “Rain rendering for evaluating and improving robustness to bad weather,” *International Journal of Computer Vision*, pp. 1–20, 9 2020. [Online]. Available: <https://link.springer.com/article/10.1007/s11263-020-01366-3>
- [55] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liang, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *CVPR*, 2020.
- [56] S. R. Richter, H. Alhaija, and V. Koltun, “Enhancing photorealism enhancement,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 02, pp. 1700–1715, feb 2023.
- [57] A. Xiao, J. Huang, D. Guan, F. Zhan, and S. Lu, “Synlidar: Learning from synthetic lidar sequential point cloud for semantic segmentation,” *arXiv preprint arXiv:2107.05399*, 2021.
- [58] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences,” in *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.
- [59] G. Eskandar, M. Abdelsamad, K. Armanious, and B. Yang, “Usis: Unsupervised semantic image synthesis,” *Computers & Graphics*, vol. 111, pp. 14–23, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0097849323000018>
- [60] G. Eskandar, D. Guo, K. Guirgis, and B. Yang, “Towards pragmatic semantic image synthesis for urban scenes,” 2023.
- [61] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, “The mapillary vistas dataset for semantic understanding of street scenes,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000–5009.
- [62] A. Barrera, J. Beltrán, C. Guindel, J. A. Iglesias, and F. García, “Cycle and semantic consistent adversarial domain adaptation for reducing simulation-to-real domain shift in lidar bird’s eye view,” in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE Press, 2021, p. 3081–3086. [Online]. Available: <https://doi.org/10.1109/ITSC48978.2021.9564553>
- [63] H. Haghghi, M. Dianati, K. Debattista, and V. Donzella, “Contrastive learning-based framework for sim-to-real mapping of lidar point clouds in autonomous driving systems,” 2023.
- [64] Y. Chen, F. Rong, S. Duggal, S. Wang, X. Yan, S. Manivasagam, S. Xue, E. Yumer, and R. Urtasun, “Geosim: Realistic video simulation via geometry-aware composition for self-driving,” *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7226–7236, 2021.
- [65] M.-F. Chang, J. W. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, and J. Hays, “Argoverse: 3d tracking and forecasting with rich maps,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [66] S. W. Kim, J. Phlion, A. Torralba, and S. Fidler, “DriveGAN: Towards a Controllable High-Quality Neural Simulation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021.
- [67] S. Zhao, Y. Wang, B. Li, B. Wu, Y. Gao, P. Xu, T. Darrell, and K. Keutzer, “epointda: An end-to-end simulation-to-real domain adaptation framework for lidar point cloud segmentation,” in *AAAI Conference on Artificial Intelligence*, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:221534728>
- [68] V. Zyrianov, X. Zhu, and S. Wang, “Learning to generate realistic lidar point cloud,” in *ECCV*, 2022.
- [69] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d,” *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.
- [70] M. Park, J. Yun, S. Choi, and J. Choo, “Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis,” 2023.
- [71] K. Nakashima and R. Kurazume, “Lidar data synthesis with denoising diffusion probabilistic models,” in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2024, pp. 14 724–14 731.
- [72] Q. Hu, Z. Zhang, and W. Hu, “Rangeldm: Fast realistic lidar point cloud generation,” 2024.
- [73] V. Zyrianov, H. Che, Z. Liu, and S. Wang, “Lidardm: Generative lidar simulation in a generated world,” 2024.
- [74] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, “Semantic image synthesis via diffusion models,” 2022.
- [75] K. Chen, E. Xie, Z. Chen, Y. Wang, L. Hong, Z. Li, and D.-Y. Yeung, “Geodiffusion: Text-prompted geometric control for object detection data generation,” 2024.
- [76] K. Yang, E. Ma, J. Peng, Q. Guo, D. Lin, and K. Yu, “Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout,” 2023.
- [77] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, “MagicDrive: Street view generation with diverse 3d geometry control,” in *International Conference on Learning Representations*, 2024.
- [78] T. Loiseau, T.-H. Vu, M. Chen, P. Pérez, and M. Cord, “Reliability in semantic segmentation: Can we use synthetic data?” 2023.
- [79] J. Su, S. Gu, Y. Duan, X. Chen, and J. Luo, “Text2street: Controllable text-to-image generation for street views,” 2024.
- [80] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, “Drivedreamer: Towards real-world-driven world models for autonomous driving,” 2023.
- [81] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, “Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving,” 2023.
- [82] X. Li, Y. Zhang, and X. Ye, “Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model,” 2023.
- [83] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, “Panacea: Panoramic and controllable video generation for autonomous driving,” 2023.
- [84] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, “Adriver-i: A general world model for autonomous driving,” 2023.
- [85] J. Lu, Z. Huang, J. Zhang, Z. Yang, and L. Zhang, “Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation,” *arXiv preprint arXiv:2312.02934*, 2023.
- [86] B. Huang, Y. Wen, Y. Zhao, Y. Hu, Y. Liu, F. Jia, W. Mao, T. Wang, C. Zhang, C. W. Chen, Z. Chen, and X. Zhang, “Subjectdrive: Scaling generative data in autonomous driving via subject control,” 2024.
- [87] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, “Drivedreamer-2: Llm-enhanced world models for diverse driving video generation,” *arXiv preprint arXiv:2403.06845*, 2024.
- [88] H. Ran, V. Guizilini, and Y. Wang, “Towards realistic scene generation with lidar diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [89] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” 2021.

- [90] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” 2023.
- [91] Q. Chen and V. Koltun, “Photographic image synthesis with cascaded refinement networks,” *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1520–1529, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8191987>
- [92] P. Vacek, O. Jašek, K. Zimmermann, and T. Svoboda, “Learning to predict lidar intensities,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 4, pp. 3556–3564, 2022.
- [93] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun, “Lidarsim: Realistic lidar simulation by leveraging the real world,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 167–11 176.
- [94] A. Hu, F. Cotter, N. Mohan, C. Gurau, and A. Kendall, “Probabilistic future prediction for video scene understanding,” Berlin, Heidelberg: Springer-Verlag, 2020, p. 767–785. [Online]. Available: https://doi.org/10.1007/978-3-030-58517-4_45
- [95] B. Guillard, S. Venmrala, J. K. Gupta, O. Miksik, V. Vineet, P. Fua, and A. Kapoor, “Learning to simulate realistic lidars,” pp. 8173–8180, 9 2022. [Online]. Available: <https://arxiv.org/abs/2209.10986v1>
- [96] Z. Li, L. Li, and J. Zhu, “Read: Large-scale neural scene rendering for autonomous driving,” in *AAAI*, 2023.
- [97] A. Ligocki, A. Jelinek, and L. Zalud, “Brno urban dataset-the new data for self-driving agents and mapping tasks,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3284–3290.
- [98] Y. Xiong, W.-C. Ma, J. Wang, and R. Urtasun, “Learning compact representations for lidar completion and generation,” in *CVPR*, 2023.
- [99] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, “Pandaset: Advanced sensor suite dataset for autonomous driving,” in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE Press, 2021, p. 3095–3101. [Online]. Available: <https://doi.org/10.1109/ITSC48978.2021.9565009>
- [100] D. Bogdol, Y. Yang, and J. M. Zöllner, “Muvo: A multimodal generative world model for autonomous driving with geometric representations,” 2023.
- [101] H. Haghighi, M. Dianati, V. Donzella, and K. Debattista, “Accelerating stereo image simulation for automotive applications using neural stereo super resolution,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12 627–12 636, 2023.
- [102] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, “World-dreamer: Towards general world models for video generation via predicting masked tokens,” *arXiv preprint arXiv:2401.09985*, 2024.
- [103] H. Haghighi, A. Samadi, M. Dianati, V. Donzella, and K. Debattista, “Taming transformers for realistic lidar point cloud generation,” 2024.
- [104] A. Swerdlow, R. Xu, and B. Zhou, “Street-view image generation from a bird’s-eye view layout,” *IEEE Robotics and Automation Letters*, vol. Preprint, 2024.
- [105] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. Kaesemodel Pontes, D. Ramanan, P. Carr, and J. Hays, “Argoverse 2: Next generation datasets for self-driving perception and forecasting,” in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. J. Vanschoren and S. Yeung, Eds., vol. 1, 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4734ba6f3de83d861c3176a6273cac6d-Paper-round2.pdf
- [106] A. van den Oord, O. Vinyals, and k. kavukcuoglu, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
- [107] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [108] J. Ost, F. Mannan, N. Thurey, J. Knodt, and F. Heide, “Neural scene graphs for dynamic scenes,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 2856–2865.
- [109] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. Srinivasan, J. T. Barron, and H. Kretzschmar, “Block-NeRF: Scalable large scene neural view synthesis,” *arXiv*, 2022.
- [110] C. Wu, J. Sun, Z. Shen, and L. Zhang, “Mapnerf: Incorporating map priors into neural radiance fields for driving view simulation,” 2023.
- [111] K. Cheng, X. Long, W. Yin, J. Wang, Z. Wu, Y. Ma, K. Wang, X. Chen, and X. Chen, “Uc-nerf: Neural radiance field for under-calibrated multi-view cameras in autonomous driving,” *arXiv preprint arXiv:2311.16945*, 2023.
- [112] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang, “Editable scene simulation for autonomous driving via collaborative Ilm-agents,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.
- [113] J. Li, Z. Feng, Q. She, H. Ding, C. Wang, and G. H. Lee, “Mine: Towards continuous depth mpi with nerf for novel view synthesis,” in *ICCV*, 2021.
- [114] H. Turki, J. Y. Zhang, F. Ferroni, and D. Ramanan, “Suds: Scalable urban dynamic scenes,” 2023.
- [115] Y. Cabon, N. Murray, and M. Humenberger, “Virtual kitti 2,” 2020.
- [116] J. Guo, N. Deng, X. Li, Y. Bai, B. Shi, C. Wang, C. Ding, D. Wang, and Y. Li, “Streetsurf: Extending multi-view implicit surface reconstruction to street views,” *arXiv preprint arXiv:2306.04988*, 2023.
- [117] Z. Wu, T. Liu, L. Luo, Z. Zhong, J. Chen, H. Xiao, C. Hou, H. Lou, Y. Chen, R. Yang, Y. Huang, X. Ye, Z. Yan, Y. Shi, Y. Liao, and H. Zhao, “Mars: An instance-aware, modular and realistic simulator for autonomous driving,” *CICAI*, 2023.
- [118] J. Yang, B. Ivanovic, O. Litany, X. Weng, S. W. Kim, B. Li, T. Che, D. Xu, S. Fidler, M. Pavone, and Y. Wang, “Emernerf: Emergent spatial-temporal scene decomposition via self-supervision,” *arXiv preprint arXiv:2311.02077*, 2023.
- [119] Z. Li, C. Wu, L. Zhang, and J. Zhu, “Dgnr: Density-guided neural point rendering of large driving scenes,” 2023.
- [120] A. Pun, G. Sun, J. Wang, Y. Chen, Z. Yang, S. Manivasagam, W.-C. Ma, and R. Urtasun, “Neural lighting simulation for urban scenes,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=mcx8IGneYw>
- [121] J. Zhang, F. Zhang, S. Kuang, and L. Zhang, “Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.
- [122] T. Tao, L. Gao, G. Wang, Y. Lao, P. Chen, Z. hengshuang, D. Hao, X. Liang, M. Salzmann, and K. Yu, “Lidar-nerf: Novel lidar view synthesis via neural radiance fields,” *arXiv preprint arXiv:2304.10406*, 2023.
- [123] S. Huang, Z. Gojcic, Z. Wang, F. Williams, Y. Kasten, S. Fidler, K. Schindler, and O. Litany, “Neural lidar fields for novel view synthesis,” 2023.
- [124] Z. Zheng, F. Lu, W. Xue, G. Chen, and C. Jiang, “Lidar4d: Dynamic neural fields for novel space-time view lidar synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [125] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, “Unisim: A neural closed-loop sensor simulator,” in *CVPR*, 2023.
- [126] A. Tonderski, C. Lindström, G. Hess, W. Ljungbergh, L. Svensson, and C. Petersson, “Neurad: Neural rendering for autonomous driving,” *arXiv preprint arXiv:2311.15260*, 2023.
- [127] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, and C. Petersson, “Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 178–20 188.
- [128] T. Tang, G. Wang, Y. Lao, P. Chen, J. Liu, L. Lin, K. Yu, and X. Liang, “Alignmif: Geometry-aligned multimodal implicit field for lidar-camera joint synthesis,” *arXiv preprint arXiv:2402.17483*, 2024.
- [129] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, Jul. 2022. [Online]. Available: <https://doi.org/10.1145/3528223.3530127>
- [130] J. Ost, I. Laradjii, A. Newell, Y. Bahat, and F. Heide, “Neural point light fields,” 2022.
- [131] M.-F. Chang, A. Sharma, M. Kaess, and S. Lucey, “Neural radiance fields with lidar maps,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17 868–17 877.
- [132] F. Lu, Y. Xu, G. Chen, H. Li, K.-Y. Lin, and C. Jiang, “Urban radiance field representation with deformable neural mesh primitives,” *ICCV*, 2023.
- [133] Y. Chen, C. Gu, J. Jiang, X. Zhu, and L. Zhang, “Periodic vibration gaussian: Dynamic urban scene reconstruction and real-time rendering,” *arXiv:2311.18561*, 2023.

- [134] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians for modeling dynamic urban scenes," 2023.
- [135] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," 2024.
- [136] H. Zhou, J. Shao, L. Xu, D. Bai, W. Qiu, B. Liu, Y. Wang, A. Geiger, and Y. Liao, "Hugs: Holistic urban 3d scene understanding via gaussian splatting," 2024.
- [137] Z. Yu, H. Wang, J. Yang, H. Wang, Z. Xie, Y. Cai, J. Cao, Z. Ji, and M. Sun, "Sgd: Street view synthesis with gaussian splatting and diffusion prior," 2024.
- [138] C. Peng, C. Zhang, Y. Wang, C. Xu, Y. Xie, W. Zheng, K. Keutzer, M. Tomizuka, and W. Zhan, "Desire-gs: 4d street gaussians for static-dynamic decomposition and surface reconstruction for urban driving scenes," 2024. [Online]. Available: <https://arxiv.org/abs/2411.11921>
- [139] W. L. Oberkampf and T. G. Trucano, "Verification and validation benchmarks," *Nuclear Engineering and Design*, vol. 238, no. 3, pp. 716–743, 2008, benchmarking of CFD Codes for Application to Nuclear Reactor Safety. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0029549307003548>
- [140] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.
- [141] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2463–2471, 2016. [Online]. Available: <https://api.semanticscholar.org/CorpusID:6746759>
- [142] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:4766599>
- [143] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Neural Information Processing Systems*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:326772>
- [144] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=rIUOzWCW>
- [145] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *ArXiv*, vol. abs/1710.10196, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3568073>
- [146] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," 2023. [Online]. Available: <https://arxiv.org/abs/2310.04378>
- [147] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Comput. Sci. Rev.*, vol. 37, p. 100270, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:198967636>



Hamed Haghghi is a PhD candidate with the Warwick Manufacturing Group (WMG) at the University of Warwick, UK. He received a B.Sc. (2016) in Software Engineering from the Isfahan University of Technology (Isfahan, Iran) and an M.Sc. (2019) in Artificial Intelligence from the University of Tehran (Tehran, Iran). His research interests include machine learning, computer vision, computer graphics, and autonomous vehicles.



Xiaomeng Wang Dr Xiaomeng Wang received a B.S. degree in Communication Engineering and an MSc degree in Information Engineering from Communication University of China in 2010 and 2013 respectively, and a PhD degree in Computer Science from the Computer Vision Laboratory at the University of Nottingham in 2018. Xiaomeng has worked as a research associate at the Graphics & Interaction Group in the Department of Computer Science and Technology, University of Cambridge, from 2017 to 2019. She is currently a research fellow in the Intelligent Vehicle Group at WMG, University of Warwick. Her main research interests involve computer vision, machine learning, and their applications.



Hao Jing Dr Hao Jing is currently a Senior Research Fellow in the Intelligent Vehicle Group at WMG, University of Warwick, with a research focus on high-performance integrated and cooperative vehicle positioning and navigation solutions in challenging environments. Dr Jing previously completed her PhD at the University of Nottingham on the topic of collaborative indoor positioning. She has worked on several projects that focus on achieving reliable and robust navigation performance for Connected and Autonomous Vehicles (CAV), pedestrians and mobile mapping systems in various environments and scenarios, based on solutions that make use of GNSS, Lidar, IMU and wireless signals.



Mehrdad Dianati Professor Mehrdad Dianati leads Networked Intelligent Systems (Cooperative Autonomy) research at Warwick Manufacturing Group (WMG), University of Warwick. He has over 28 years of combined industrial and academic experience, with 20 years in leadership roles in multi-disciplinary collaborative R&D projects, in close collaboration with the Automotive and ICT industries. He is also the Co-Director of Warwick's Centre for Doctoral Training on Future Mobility Technologies, training doctoral researchers in the areas of intelligent and electrified mobility systems in collaboration with the experts in the field of electrification from the Department of Engineering of the University of Warwick. In the past, he served as an associate editor for the *IEEE Transactions on Vehicular Technology* and several other international journals, including *IET Communications*. Currently, he is the Field Chief Editor of *Frontiers in Future Transportation*. His academic experience includes over 25 years of teaching undergraduate and post-graduate level courses and supervision of research students. He currently leads a post-graduate course in Machine Intelligence.