StreetSyn: A Full Radiance Field Solution for Street and Vehicle Free-View Synthesis

Shenhao Zhu[1], Li Wang[1], Xun Cao[1], Ruigang Yang[2], Xinxin Zuo[3], and Hao Zhu[1]

[1] Nanjing University, Nanjing & Suzhou, China
[2] Inceptio, Shanghai, China
[3] University of Alberta, Edmonton, Canada

Abstract. Starting from sparse views of real-captured scenes and a synthetic dataset of 3D vehicles, we aim to synthesize photo-realistic street views with moving vehicles, editable illumination, and controllable viewpoints which is a significant task for autonomous driving simulation. The problem is very challenging as only sparse views are available for recovering such a complex street environment. In this paper, we propose a full radiance field scheme for free-view synthesis of street scenes and vehicles. Benefiting from the scheme that both the scene and the vehicle are represented as radiance fields, illumination can be directly extracted from the real-captured scenes and transferred to the synthesized vehicle. The ambient illumination is modeled as a mixture of Spherical Gaussians(SGs) with different frequencies, which turns out to be effective in recovering the low-frequency sky illumination and high-frequency sun illumination. Experiments show that our model can synthesize street view and vehicle images in free views, and significantly outperforms previous works in photo-realism and lighting modeling accuracy.

Keywords: Scene synthesis, Driving simulation, Neural Radiance Field

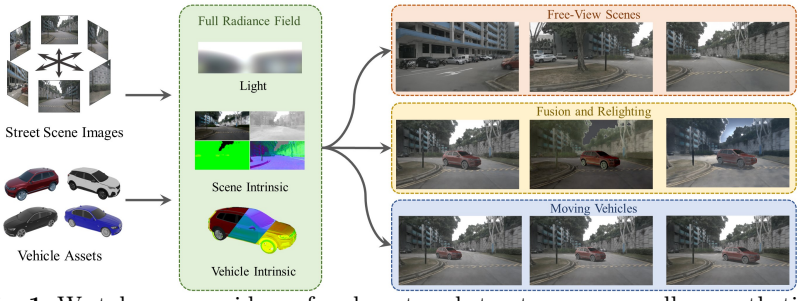


Fig. 1. We take sparse videos of real-captured street scenes as well as synthetic vehicles as input. Full neural radiance fields of both geometric and photometric properties such as lighting, materials are optimized for both the background scenes and synthetic vehicles to support flexible street scene simulation, such as free-view rendering, relighting, vehicle composition and manipulation in 3D.

1 Introduction

Humans can envision unseen scenarios in their minds, significantly enhancing the effectiveness of learning. Drawing inspiration from this capacity, researchers in the field of autonomous driving have focused on creating virtual environments with synthesized vehicles to train perception models. Extensive studies demonstrate that this approach is crucial in addressing challenges related to 3D visual-based decision-making problems [21, 36].

The traditional 3D simulation involves the utilization of Computer-Aided Design (CAD) models. Nevertheless, generating extensive and varied street scenes with purely CAD models proves highly inefficient and costly in practical terms. On the other hand, there are extensive street view datasets of real-captured scenes, which contain extremely diverse street scenes under real-world lighting information. Thus, in this paper, we propose a hybrid solution that models the street scenes with real-captured sparse videos and composites of synthetic 3D vehicles to synthesize photo-realistic free-view rendering in diverse scenarios.

Pioneering works try to synthesize street views by modeling vehicles as 3D objects and representing the background scenes as static images [5, 20, 32]. These methods exhibit limited control over the overall appearance of both scenes and vehicles, and the viewpoint remains fixed. Now benefiting from the emergence of neural radiance field (NeRF) [24] for a 3D scene representation, we can model background scenes and integrated vehicles separately and then composite them into a single radiance field to produce high rendering quality in free-view. Some recent works [2, 33, 35] exploited NeRF representation for street scene rendering and proposed to reconstruct scene geometry and recover intrinsic properties of the scene for relighting purpose. Nevertheless, the rendering quality is significantly constrained by the simplified lighting and material modeling employed to depict intricate street scenes.

As shown in Fig. 1, in this paper, we propose a full neural radiation field solution to build a street scene simulator, which supports free-view rendering, 3D vehicle composition, and relighting. The key idea is that by modeling both real-captured street view scenes and virtual vehicles as neural radiance fields we can integrate them in 3D space by considering both geometric constraints and the illumination effects.

First, to attain photo-realistic rendering and relighting of integrated background scenes and virtual vehicles, we put lots of effort into decomposing the intrinsic scene properties. Basically, we propose a novel scheme to effectively decompose the environmental lighting while optimizing the NeRF for captured street scenes. Specifically, we model the ambient illumination as a mixture of SGs with different frequencies, which turns out to be effective in recovering the low-frequency sky illumination and high-frequency sun illumination in a real-world environment. Second, for the virtual vehicles, we build up a neural radiance field library with their decomposed intrinsic properties as well. By leveraging the radiance fields of both background scenes and virtual vehicles, along with their intrinsic properties, we can execute scene composition in accordance with spatially varying rendering equations. Finally, a FusionNet is introduced to further

eliminate the synthetic-to-real gap after the composition of the inserted vehicles and the street scene, producing photo-realistic renderings.

Our contribution can be summarized as:

- To the best of our knowledge, our method is the first full radiance field solution for street and vehicle free-view synthesis for autonomous driving simulation, which supports free-view rendering of the scene and greatly expands the simulation scene.
- To achieve realistic relighting and composition, we propose a method to decompose intrinsic properties of both street scenes and virtual vehicles while optimizing NeRF.
- Inverse rendering and a generative refinement are introduced to eliminate the bias between real-captured data and synthetic data, yielding photo-realistic fusing images.

2 Related Work

We first introduce the previous works of simulation for autonomous driving, then review the relighting and composition for synthesized objects, which are two key modules in generating photo-realistic scenes.

Driving view simulation: Simulating driving views has been a key technique in training an autonomous driving system and is catching wide attention. The traditional simulator [7, 27] leverages manually designed 3D models to render street scenes with moving vehicles. These renderings contain a synthetic-to-real gap, and a limited amount of 3D models makes it very difficult to build large-scale and diverse scenes. In the follow-up research, the performance boost of 2D conditioned generation models [4,15,26,31] provides an alternative to synthesize novel scenes with editable contents in 2D generated images [8,21]. These methods struggle in simulating 3D-related properties like rotations, shadows, and occlusions. Very recently, Neural Radiance Field (NeRF) [24] was introduced for its high-quality rendering and free-view synthesis performance [1, 5, 34, 37]. These works represent vehicles with a NeRF and then fuse the renderings into scene images. Though 3D properties for vehicles are learned, the rendering views are limited as the scenes are still in a 2D representation. More importantly, the vehicles and scenes are represented in a radiance field and an image separately, so the lighting conditions cannot be accurately estimated and transferred, leading to a less realistic synthesis in complex and diverse lighting situations. In contrast, our method is a full radiance field solution merging street scenes and vehicles, going further in synthesizing relightable and free-view renderings.

Merging objects and scenes: Merging synthetic objects (vehicles) into reconstructed outdoor street scenes is a valuable task and of great significance to solving the long-tail scenario problem of autonomous driving. To obtain photo-realistic fusion results, some works [9, 12–14, 19] start from the background environmental images, trying to collect information from an outdoor background image to obtain the surrounding environmental lighting conditions, and represent the scene lighting as spherical harmonic coefficients or HDR maps of sun

and sky. However, these methods cannot cover the impact of objects in the scene on lighting and the effect changes that occur after foreground objects are inserted. To solve this problem, some methods [29, 38] decompose lighting into global and local parts for prediction respectively. However, due to the limitation that the background is a 2D image, the predicted illumination cannot guarantee the spatial consistency of the same scene.

After NeRF was proposed, some methods [22, 28, 33] used intrinsic decomposition and inverse rendering images to optimize scene geometry while estimating illumination to ensure spatial consistency. For example, NeRF-OSR [28] proposes a method that combines spherical harmonic coefficient illumination and diffuse reflection effects, but it cannot restore the high-frequency highlight effect. Based on the characteristics of the natural environment, we choose SG mixtures to represent environmental lighting, use a series of regularization terms to constrain and optimize the decomposition of the scene's intrinsic attributes, and use an inverse renderer to restore the color and shadow effects of foreground objects to the greatest extent. In addition, we use FusionNet to further enhance the realism after fusion, thereby proposing a complete solution to this problem.

3 Method

3.1 Overview

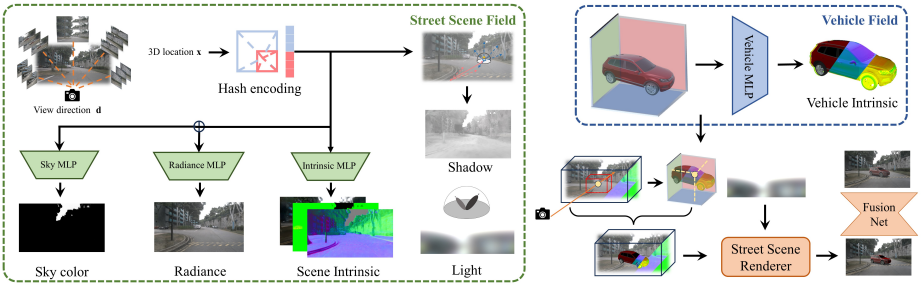


Fig. 2. Overall Pipeline. Our approach enables moving foreground objects to be inserted into real-world scenes. First, two neural radiance fields are generated for street scenes and vehicles separately, with the intrinsic properties decomposed. Then the two fields are fused to synthesize a lighting-uniform street scene with vehicles inserted.

As shown in Fig. 2, our approach enables moving foreground objects to be inserted into real-world scenes and is capable of generating a high-quality video of the composition in novel views. For the background scene, we take as input multiple frames of images extracted from a street scene video, the camera poses can be obtained from the IMU/GPS information, denoted as $\{I_i, C_i\}_{i=1}^{N}$, where $I_i \in \mathbb{R}^{H*W*3}$ is an image, $C_i \in SE(3)$ is its camera pose and N is the number of images. A neural radiance field for a street scene is built to decompose the intrinsic scene properties (Sec 3.2).

We introduce a mixture of Spherical Gaussians(SGs) [30] with different frequencies to model ambient illuminations, which is sufficient to cover the low-frequency sky illumination and relatively high-frequency sun illumination in a real-world environment. A Street Scene Renderer is leveraged to jointly optimize the parameters of the neural field and lighting conditions in the inference phase (Sec 3.3). During training, we model the sky and the scene separately and design their respective loss functions. A set of regularization terms optimizes this highly underdetermined problem (Sec 3.4). For foreground objects, we take advantage of a virtual vehicle library and build a network to decompose the intrinsic properties of these vehicles. The vehicle information can be queried through the library, and the vehicle can be directly inserted into the street scene with a modified rendering equation. FusionNet is used to handle inconsistent color saturation and harsh boundaries and to eliminate the synthetic-to-real gap (Sec 3.5). It is worth noting that our approach represents both foreground objects and background scenes in a unified field with their intrinsic properties decomposed, which enables the scene to be rendered in free views.

3.2 NeRF for Street Scenes

NeRF [24] represents a scene with a neural field $F : (\boldsymbol{x}, \boldsymbol{d}) \mapsto (\sigma, \boldsymbol{c})$ which maps 3D location $\boldsymbol{x}$ and view direction $\boldsymbol{d}$ to corresponding density $\sigma \in \mathbb{R}$ and color $\boldsymbol{c} \in \mathbb{R}^3$. To model a large-scale street scene, hash-based NeRF representation [25] is leveraged to improve the efficiency of training and inference.

Intrinsic property decomposition: A neural intrinsic field $F_\phi : \boldsymbol{x} \mapsto (\sigma, \boldsymbol{n}, \boldsymbol{k})$ is learned to decompose the intrinsic properties related to spatial location. Specifically, we encode input 3D location $\boldsymbol{x}$ with a hash table, then use three separated modules to decode the hash feature: the geometry module $\sigma = F_{geo}(\boldsymbol{x}; \theta_g)$ outputs the density $\sigma \in \mathbb{R}$; the normal module $\boldsymbol{n} = F_{norm}(\boldsymbol{x}; \theta_n)$ outputs the surface normal $\boldsymbol{n} \in \mathbb{R}^3$; and the appearance module $\boldsymbol{k} = (\boldsymbol{k}_d, \boldsymbol{k}_s) = F_{app}(\boldsymbol{x}; \theta_a)$ outputs the material properties of the surface point.

Virtual vehicle library: To enhance the diversity of our model's synthetic scenes, we have collected a series of virtual vehicle models and established a Virtual Vehicle Library to complement our street scene field. This enables the synthesis of various foreground vehicles, enhancing the diversity of the scenes. Each vehicle is represented as an implicit representation with triplanes and a small MLP decoder. we obtain the intrinsic decomposition of the vehicles, maintaining consistency with the scene, yielding a vehicle module $(\sigma_v, \boldsymbol{k}_v) = F_v(\boldsymbol{x}; \theta_v)$. This allows for direct integration of the vehicles into the scene at intrinsic level.

3.3 Street Scene Renderer

The rendering equation: The rendering equation [17] describes how light propagates in a 3D scene and determines the irradiance observed at a particular point. It provides a common formulation for rendering in computer graphics:

$$L_o(\boldsymbol{p}, \boldsymbol{w}_o) = \int_\Omega f_r(\boldsymbol{p}; \boldsymbol{w}_i, \boldsymbol{w}_o) L_i(\boldsymbol{p}, \boldsymbol{w}_i)(\boldsymbol{w}_i \cdot \boldsymbol{n}) d\boldsymbol{w}_i \tag{1}$$

Following the rendering equation, given a camera ray $r(t) = \boldsymbol{o} + t\boldsymbol{d}$, we first compute the location of $\boldsymbol{p}$ by the volume render depth $\hat{D}$ as $\boldsymbol{p} = \boldsymbol{o} + \hat{D}\boldsymbol{d}$, then use an inverse render workflow to synthesize photorealistic street scenes.

Lighting representation: As the rendering equation is an integral equation without an analytical solution, existing methods either simplify the ambient lighting [22,28] or calculate it through expensive Monte Carlo integration [10,33], which reduces the fidelity and efficiency of rendering.

To address the problem, We represent the ambient lighting as a sum of spherical Gaussian lobes (SGs) [30]:

$$L_i(\boldsymbol{\omega_i}) = \sum_{k=1}^{M} G(\boldsymbol{\omega_i}; \boldsymbol{\phi}, \lambda, \boldsymbol{\mu}) \tag{2}$$

where $\boldsymbol{\phi} \in \mathbb{R}^3$ is the direction, $\lambda \in \mathbb{R}_+$ is the sharpness and $\boldsymbol{\mu} \in \mathbb{R}^3_+$ is the amplitude, for a particular SG lobe. We regard the SGs as a part of model parameters and jointly optimize them during the training stage.

Metal/roughness workflow: Street scenes encompass a wide variety of objects, each with distinct materials. In such scenes, many objects, such as bricks, roads, and plants, typically have a high roughness, resulting in diffuse reflection effects. However, for vehicles, the car paint often exhibits metallic properties and characteristics like lacquer, which cannot be accurately simulated by considering roughness alone. Therefore, to precisely decompose and render the intrinsic properties of street scenes, we introduce the metal/roughness workflow to model the materials in street scenes.

In our workflow, the material properties $\boldsymbol{k}_d$ and $\boldsymbol{k}_s$ control the diffuse appearance and the specular appearance respectively. $\boldsymbol{k}_d \in \mathbb{R}^3$ represents the base color of the object's surface. In our workflow, $\boldsymbol{k}_s$ is composed of roughness α and metalness m from the material properties.

Based on the aforementioned material properties, we define the BRDF in the Metal/roughness workflow as the sum of the diffuse component and the specular component: $f_r = f_d + f_s$. The diffuse part is determined by the base color $\boldsymbol{k}_d$ and the metalness m:$f_d = \frac{\boldsymbol{k}_d}{\pi} * (1 - m)$. The metalness here is used to control the intensity of the diffuse reflection. The specular component of BRDF consists of normal distribution function, Fresnel term and shadowing term. Metalness m serves here to align the color of the specular highlights with the base color, thereby enhancing the metallic sensation. Similar to [41], We represent the specular component of BRDF as a single SG, where $\boldsymbol{h}$ is exactly halfway between the light direction vector and the view direction vector. :

$$f_s = G(\boldsymbol{h}; \boldsymbol{n}, \frac{1}{2\alpha^2(\boldsymbol{h} \cdot \boldsymbol{\omega_o})}, \frac{\mathcal{M}}{\pi\alpha^2}) \tag{3}$$

As both lighting and BRDF are represented by SGs, we can first calculate the lighting and BRDF, then obtain the emitted irradiance C_{render} by closed-form hemispherical integration of the SGs [23].

Shadow map: Owing to the intricate nature of street scenes, the presence of various objects obstructs light, leading to the formation of shadows throughout

the scene. We compute a shadow map by uniformly sampling N_l incident rays on the upper hemisphere, the visibility V of an incident ray with direction $\boldsymbol{l}_k$ at surface point $\boldsymbol{p}$ can be denoted as the volume opacity of the ray: $r(t) = \boldsymbol{p} + t\boldsymbol{l}_k$. The ray is traced through the scene to obtain the visibility by volume rendering:

$$V = \exp(-\sum_i \sigma_i(t_i - t_{i-1})), \tag{4}$$

Then, we generate the shadow using a ratio between the occluded and unoccluded irradiance.

$$S(\boldsymbol{p}) = \frac{\sum_{k=1}^{N_l} f_r(\boldsymbol{p}; \boldsymbol{\omega_o}, \boldsymbol{l}_k) L_i(\boldsymbol{p}; \boldsymbol{l}_k) V(\boldsymbol{p}; \boldsymbol{l}_k)(\boldsymbol{l}_k \cdot \boldsymbol{n})}{\sum_{k=1}^{N_l} f_r(\boldsymbol{p}; \boldsymbol{\omega_o}, \boldsymbol{l}_k) L_i(\boldsymbol{p}; \boldsymbol{l}_k)(\boldsymbol{l}_k \cdot \boldsymbol{n})} \tag{5}$$

A shadow map can be added to the scene by a pixel-wise product: $L'_o = S \bigodot L_o$

3.4 Optimization and Regularization

The sparse training image perspective and unknown lighting conditions make the scene reconstruction an extremely ill-posed problem. Therefore, several data augmentation and regularization terms are leveraged to constrain the problem.
Rendering loss: We use the consistency between the predicted color and ground-truth color from input images as our main supervision, which is formulated as:

$$\mathcal{L}_{\text{render}} = \sum_{r \in \mathcal{R}} |C_{\text{gt}} - C_{\text{render}}|^2 \tag{6}$$

where C_{render} is the RGB value calculated by the rendering pass of each camera ray $r \in \mathcal{R}$, and C_{gt} is the corresponding ground-truth color. An end-to-end training scheme is achieved with attribute decomposition and lighting parameters jointly optimized.
Sky mask loss: We use a binary cross-entropy loss term between the volume rendered alpha channel mask M_α and sky mask M_textgt, which is obtained from an off-the-shelf semantic segmentation network [16]. The sky mask loss term $\mathcal{L}_{\text{skymask}}$ is formulated as:

$$\mathcal{L}_{\text{skymask}} = \sum_{r \in \mathcal{R}} \text{BCE}(M_\alpha, M_{\text{gt}}) \tag{7}$$

Sky modeling loss: Because performing physically-based rendering on the pixels of the sky is meaningless, we used an off-the-shelf semantic segmentation network [16] to mask out the sky pixels. We first use a binary cross-entropy to find sky pixels, then the sky network can give the corresponding sky color according to the viewing direction. An MSE loss is used to evaluate the difference between the predicted sky color and the ground-truth sky color, formulated as:

$$\mathcal{L}_{\text{sky}} = \sum_{r \in \text{sky}} |C_{\text{gt}} - C_{\text{sky}}|^2 \tag{8}$$

Normal loss: Normal extracted from the density field $\hat{\boldsymbol{n}} = -\frac{\nabla\sigma(\boldsymbol{x})}{|\nabla\sigma(\boldsymbol{x})|}$ is used as a supervisory term for normal predicted from MLP. Additionally, we use an off-the-shelf normal estimator [18] to provide a reference normal $\boldsymbol{n}_r$, as in:

$$\mathcal{L}_{\text{norm.}} = \sum_{r\in\mathcal{R}} \left(|\boldsymbol{n}_r - \boldsymbol{n}|^2 + |\hat{\boldsymbol{n}} - \boldsymbol{n}|^2\right) \tag{9}$$

Base color loss: We observed that without any constraints, the shadows in the scene tend to be incorrectly merged into the base color. Noting that shadows often appear on roads, we utilize a shadow detector [6] and a segmentation network [16]. The pixels identified by the shadow detector as being in shadow, and the road pixels, are combined to define the domain $\mathcal{S}$ where constraints should be applied to the base color. Then we mask out the shadow pixels in the image and calculate the average color of each semantic segment after removing shadows to serve as the reference color $C_{\text{ref.}}$. We encourage the base color to be consistent with the pixels without shadow, which is formulated as:

$$\mathcal{L}_{\text{base.}} = \sum_{r\in\mathcal{S}} |C_{\text{ref.}} - C_{\text{base.}}|^2 \tag{10}$$

Combining all supervision and regularization items, our overall loss function is denoted as:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{render}} + \lambda_{\text{skymask}}\mathcal{L}_{\text{skymask}} \\ &+ \lambda_{\text{sky}}\mathcal{L}_{\text{sky}} + \lambda_{\text{norm.}}\mathcal{L}_{\text{norm.}} + \lambda_{\text{base.}}\mathcal{L}_{\text{base.}}\end{aligned} \tag{11}$$

In our experiments, we set $\lambda_{\text{rad.}} = \lambda_{\text{sky}} = 1$, $\lambda_{\text{skymask}} = \lambda_{\text{norm.}} = \lambda_{\text{base.}} = 0.1$, $\lambda_{\text{smooth}} = 0.01$. λ_{depth} is set to 0.1 on nuScenes [3] dataset.

3.5 Street Scene Fusion

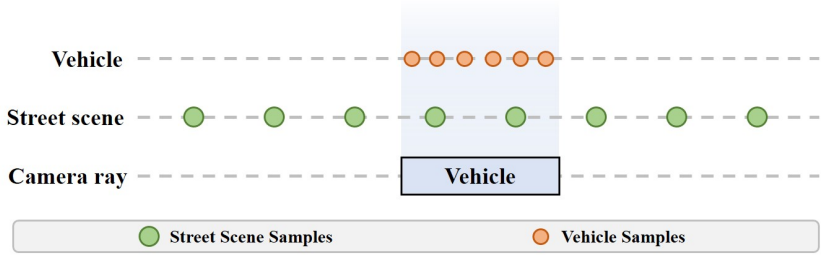


Fig. 3. We employ a hybrid sampling approach to facilitate the fusion of vehicles with street scenes. Initially, vehicles and street scenes are sampled independently, followed by sorting and combining them according to depth in the real-world coordinate.

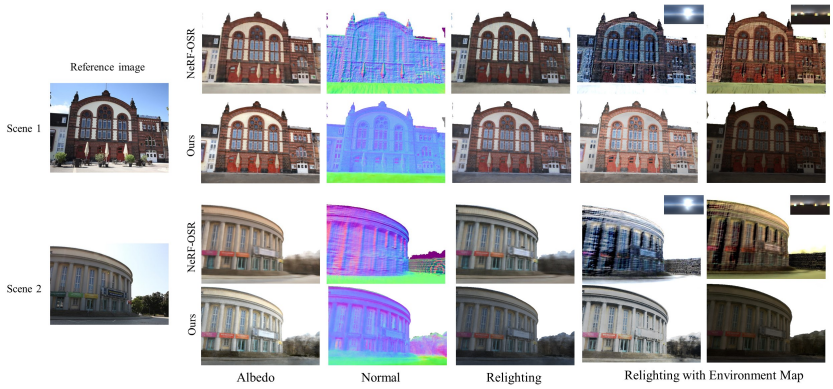


Fig. 4. Qualitative results of relighting on NeRF-OSR dataset [28]. Our method reconstructs normals with less noise and restores realistic shadows, enabling high-quality relighting results under various lighting conditions.

Intrinsic properties merging: As shown in Fig. 3, a hybrid sampling method is used to integrate vehicles and street scenes together. Beyond the original sampling points within the scene, we introduce an additional set of sampling points in the real-world coordinate. These additional points are specifically designated for the sampling of vehicles.

After the vehicle and the scene are both sampled, we sort and combine the sampling points based on their depth in the real-world coordinate, resulting in the combined sampling points $\{X\}$. Then, let $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ denote the camera ray with origin $\mathbf{o}$ and direction $\mathbf{d}$, we obtain color and intrinsic properties of the vehicle and scene fusion by performing standard volume rendering:

$$\boldsymbol{c}(\mathbf{r}) = \sum_{X_i} T_i \alpha_i \boldsymbol{c}_i, \quad T_i = \exp(-\sum_{k=1}^{i-1} \sigma_k \delta_k) \tag{12}$$

$$\boldsymbol{n}(\mathbf{r}) = \sum_{X_i} T_i \alpha_i \boldsymbol{n}_i, \quad \boldsymbol{k}(\mathbf{r}) = \sum_{X_i} T_i \alpha_i \boldsymbol{k}_i \tag{13}$$

where $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$, $\delta_i = t_{i+1} - t_i$. Similar to [34], we truncate the density of scene sampling points within the vehicle's bounding box to prevent confusion between the scene and the vehicle. We can extract the mask of the foreground vehicle by comparing the depth of the vehicle and the scene, serving as the input for our FusionNet.

Fusion network: Our Street Scene Renderer has achieved a photo-realistic composition of the appearance and shadow of foreground vehicles and street scenes. However, there are still some differences between the virtual vehicle models and the real vehicles, such as dust on the car body and fading of the car paint, etc. Therefore, we propose FusionNet to further eliminate the differences between virtual and real. We segmented vehicles from the street scene dataset and disturbed the clarity and color saturation of the vehicles to simulate the differences between virtual and real. We used a network structure similar to [39], with the

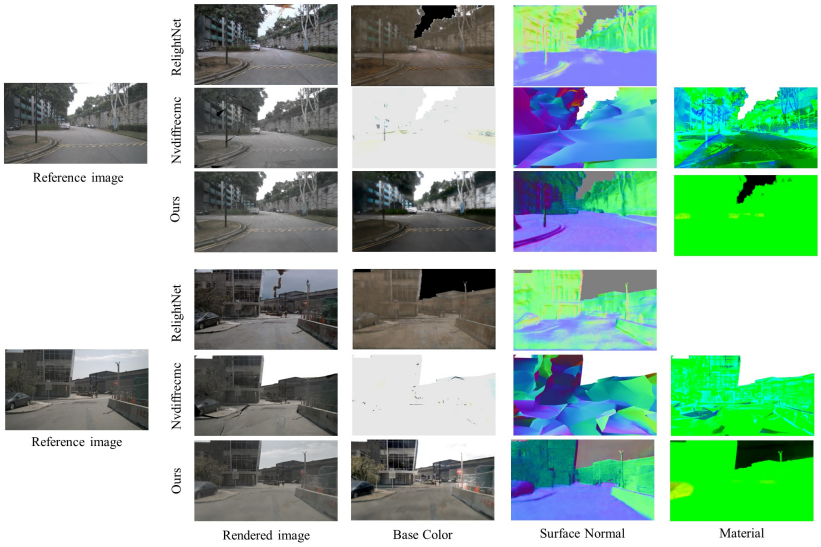


Fig. 5. Qualitative comparison of intrinsic properties decomposition. Our method reconstructs the true base color without any residual shadows. Furthermore, we achieve smoother normals and consistent materials, enhancing rendering quality.

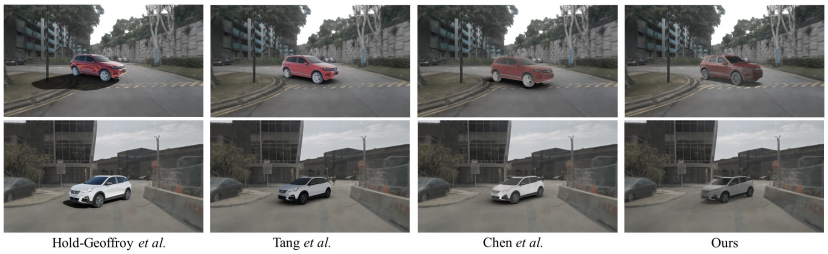


Fig. 6. Qualitative comparison of foreground vehicle fusion. Our method harmoniously blends vehicles with street scenes, accurately restoring details such as highlights and shadows on the vehicles.

addition of a vehicle segmentation mask as input, and an extra supervision item for color to ensure that the appearance of vehicles stays unchanged.

4 Experiments

4.1 Datasets and Baselines

NuScenes dataset: NuScenes dataset [3] provides comprehensive resources for autonomous driving research. To simulate an urban environment, we select data from two distinct scenes from nuScenes dataset. Each scene's data encapsulates a video recorded by six strategically placed cameras on a moving vehicle. In our experiments, we used images from all six viewpoints.

NeRF-OSR dataset: NeRF-OSR [28] dataset contains eight sites captured from 3240 viewpoints using a DSLR camera across 110 different recording ses-



Fig. 7. Fusion results when removing or replacing a certain module of our proposed fusion process for ablation study.

	Scene 1		Scene 2	
Method	PSNR↑	MSE↓	PSNR↑	MSE↓
NeRF-OSR [28]	15.86	0.026	15.97	0.025
Ours (w/o norm. reg.)	17.12	0.019	17.83	0.016
Ours (w/o shadow)	17.73	0.017	18.12	0.015
Ours	**17.95**	**0.016**	**18.56**	**0.014**

Table 1. Quantitative Evaluation on NeRF-OSR [28] dataset

sions. For each site, a 360-degree shot of the environment map was also taken. Images from each site are divided into a training set and a test set with ground-truth illumination maps. We applied our Neural Street Field model to two scenes from NeRF-OSR dataset, and obtained quantitative evaluation results.

Baseline methods: We use different baselines for different tasks for comparison. We establish Neural Street Field on NeRF-OSR dataset, and make qualitative and quantitative comparisons in terms of the quality of rendering and relighting. To decomposite intrinsic properties of scenes, we compare our method with RelightNet [40] and Nvdiffrecmc [11]. Lastly, we compare the methods of 2D illumination estimation, examining the effects of using generative network techniques in the integration of foreground vehicles and street scenes.

4.2 Rendering and Relighting Quality

Outdoor relighting quality: As shown in Fig. 4, We select two scenes from NeRF-OSR dataset and build the Neural Street Field model to conduct relighting evaluations for outdoor scenes. Our quantitative evaluation results are reported in Tab. 1. Our method outperforms previous methods in PSNR and MSE by a large margin, recovering more accurate albedo, normal, and shadow effects. We also evaluate the relighting effect on two environment maps out of the dataset. The spherical harmonics based lighting representation of NeRF-OSR cannot fully capture the global lighting condition, and its bumpy normal also causes poor relighting effects under bright lighting conditions. We comprehensively improve the reconstruction quality of various intrinsic properties, thereby

achieving much better relighting effects than NeRF-OSR. The additional ablation study is conducted by: (a) Without normal regularization: We do not use additional normal regularization terms (b) Without shadow: The shadow map is removed, which means all surface points can be illuminated by the light source. **Intrinsic decomposition quality:** Autonomous driving datasets commonly provide multi-view videos in a center-outward manner, resulting in very sparse viewing angles of the scene. Such images acquire limited 3D geometric information about the scene. Additionally, the images in the dataset commonly suffer from motion blur and inconsistent brightness, which poses significant challenges to accurately decomposing the intrinsic properties of the scene. As shown in Fig. 5, We compare the quality of intrinsic decomposition with RelightNet [40] and Nvdiffrecmc [11] RelightNet is unable to decompose material information and shows significant errors in estimating base color and normal. Nvdiffrecmc fails to reconstruct correct geometry and appearance from sparse observation angles. Compared with these two methods, our method achieves high-quality intrinsic decomposition that can be directly used for render passes, producing high-quality relighting results.

4.3 Vehicle Fusion Quality

Qualitative comparison: As shown in Fig. 6, we compare the fusion effect of our method with other methods. Hold-Geoffroy [12] first proposed a method for estimating sky illumination, but they ignored the occlusion caused by objects in the scene, leading to over bright appearance and incorrect shadows. Tang [29] considered both global and local lighting, which led to a more realistic appearance. However, due to the lack of 3D constraints, their results lacked consistency and were unable to bridge the gap between synthetic objects and real-world objects. Chen [5] tried using a generative network to modify the appearance of vehicles, but they could not restore details such as highlights and shadows. By contrast, our method synthesizes accurate highlights and shadow effects, leading to photo-realistic insertion of vehicles into the scene.

4.4 Ablation Study

To validate the effectiveness of the introduction of shadow map, Metal/roughness workflow and FusionNet, we conduct the experiments with the following settings:
• (a) Without occulusion: The shadow map is removed. All surface points are considered visible to the light source.
• (b) Without f_s: The specular component f_s of the BRDF is removed, retaining only the diffuse appearance.
• (c) Without metallic: The metalness of the vehicle is set to 0, modeling the vehicle as a completely non-metallic object.
• (d) Without specular: The metalness in specular component of f_s is set to 0, and the specular color turns white.
• (e) Without FusionNet: The FusionNet is removed from the fusion process.

• (f) Ours Full: Our method fuses a vehicle into a street scene.

The visualized results of the ablation study are shown in Fig. 7. We find that our full method achieves a photorealistic composition of the foreground vehicle and the street scene. Comparing (a) with (f), we can see that there are no shadow effects on the ground near the vehicle, and abnormal bright spots appear on the body of the car. This demonstrates the necessity of considering the shadows generated by objects within the scene and by the vehicle itself. Comparing (b) with (f), we find that the vehicle becomes a completely Lambertian surface, without any specular effects, failing to reflect the true appearance of the vehicle's material. This proves the advantages of the Metal/roughness workflow. Comparing (c) with (f), the appearance of the vehicle shifts towards white, this demonstrates the superiority of incorporating metalness into out Street Scene Renderer. Comparing (d) with (f), metalness makes the highlight display the correct color. Comparing (e) with (f), in the physically-based rendering results on virtual vehicles, there are issues of color inconsistency and over-saturation. FusionNet can eliminate the gap between virtual and real, yielding indistinguishably photorealistic results.

5 Conclusion

In this paper, we propose to synthesize photo-realistic street views with moving vehicles, editable illumination, and controllable viewpoints from sparse views of real-captured scenes and a synthetic dataset of 3D vehicles. A full radiance filed scheme is introduced for free-view synthesis of street scenes and vehicles, where illumination can be directly extracted from the real-captured scenes and transferred to the synthesized vehicle. Experiments show that our model can synthesize street view and vehicle images in free views, and significantly outperforms previous works in photo-realism and lighting modeling accuracy.

Limitations: Although our method achieves high-quality rendering of street scenes and fusion with foreground vehicles, it still has certain limitations. Our method remains reliant on some 2D priors, and additionally, we have not accounted for the effects of multiple light bounces. In our future work, we aim to extract more information from existing data to reduce our dependence on 2D priors. We plan to employ more sophisticated rendering techniques to minimize energy loss during the rendering process.

References

1. Abeysirigoonawardena, Y., Xie, K., Chen, C., Hosseini, S., Chen, R., Wang, R., Shkurti, F.: Generating transferable adversarial simulation scenarios for self-driving via neural rendering. arXiv preprint arXiv:2309.15770 (2023)
2. Agrawal, D., Xu, J., Mustikovela, S.K., Gkioulekas, I., Shrivastava, A., Chai, Y.: Nova: Novel view augmentation for neural composition of dynamic objects. In: ICCV. pp. 4288–4292 (2023)

3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV. pp. 1511–1520 (2017)
5. Chen, Y., Rong, F., Duggal, S., Wang, S., Yan, X., Manivasagam, S., Xue, S., Yumer, E., Urtasun, R.: Geosim: Realistic video simulation via geometry-aware composition for self-driving. In: CVPR. pp. 7230–7240 (2021)
6. Chen, Z., Zhu, L., Wan, L., Wang, S., Feng, W., Heng, P.A.: A multi-task mean teacher for semi-supervised shadow detection. In: CVPR (2020)
7. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
8. Fang, J., Zhou, D., Yan, F., Zhao, T., Zhang, F., Ma, Y., Wang, L., Yang, R.: Augmented lidar simulator for autonomous driving. IEEE Robotics and Automation Letters **5**(2), 1931–1938 (2020)
9. Gao, D., Li, X., Dong, Y., Peers, P., Xu, K., Tong, X.: Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. ToG **38**(4), 134–1 (2019)
10. Hasselgren, J., Hofmann, N., Munkberg, J.: Shape, light, and material decomposition from images using monte carlo rendering and denoising. NIPS **35**, 22856–22869 (2022)
11. Hasselgren, J., Hofmann, N., Munkberg, J.: Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. arXiv:2206.03380 (2022)
12. Hold-Geoffroy, Y., Athawale, A., Lalonde, J.F.: Deep sky modeling for single image outdoor lighting estimation. In: CVPR. pp. 6927–6935 (2019)
13. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: CVPR. pp. 7312–7321 (2017)
14. Hosek, L., Wilkie, A.: An analytic model for full spectral sky-dome radiance. ACM Trans. Graph. **31**(4) (jul 2012). https://doi.org/10.1145/2185520.2185591, https://doi.org/10.1145/2185520.2185591
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017)
16. Jain, J., Li, J., Chiu, M., Hassani, A., Orlov, N., Shi, H.: OneFormer: One Transformer to Rule Universal Image Segmentation (2023)
17. Kajiya, J.T.: The rendering equation. In: 13th annual conference on Computer graphics and interactive techniques. pp. 143–150 (1986)
18. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3d common corruptions and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18963–18974 (2022)
19. Lalonde, J.F., Matthews, I.: Lighting estimation in outdoor image collections. In: Proceedings of the 2014 2nd International Conference on 3D Vision - Volume 01. p. 131–138. 3DV '14, IEEE Computer Society, USA (2014). https://doi.org/10.1109/3DV.2014.112, https://doi.org/10.1109/3DV.2014.112
20. Li, L., Lian, Q., Wang, L., Ma, N., Chen, Y.C.: Lift3d: Synthesize 3d training data by lifting 2d gan to 3d generative radiance field. In: CVPR. pp. 332–341 (2023)
21. Li, W., Pan, C., Zhang, R., Ren, J., Ma, Y., Fang, J., Yan, F., Geng, Q., Huang, X., Gong, H., et al.: Aads: Augmented autonomous driving simulation using data-driven algorithms. Science Robotics **4**(28), eaaw0863 (2019)

22. Lin, Z.H., Liu, B., Chen, Y.T., Forsyth, D., Huang, J.B., Bhattad, A., Wang, S.: Urbanir: Large-scale urban scene inverse rendering from a single video. arXiv preprint arXiv:2306.09349 (2023)
23. Meder, J., Brüderlin, B.: Hemispherical gaussians for accurate light integration. In: ICCVG. pp. 3–15. Springer (2018)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
25. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ToG **41**(4), 1–15 (2022)
26. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR. pp. 2337–2346 (2019)
27. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR. pp. 3234–3243 (2016)
28. Rudnev, V., Elgharib, M., Smith, W., Liu, L., Golyanik, V., Theobalt, C.: Nerf for outdoor scene relighting. In: ECCV. pp. 615–631. Springer (2022)
29. Tang, J., Zhu, Y., Wang, H., Chan, J.H., Li, S., Shi, B.: Estimating spatially-varying lighting in urban scenes with disentangled representation. In: European Conference on Computer Vision. pp. 454–469. Springer (2022)
30. Wang, J., Ren, P., Gong, M., Snyder, J., Guo, B.: All-frequency rendering of dynamic, spatially-varying reflectance. In: SIGGRAPH Asia. pp. 1–10 (2009)
31. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR. pp. 8798–8807 (2018)
32. Wang, Z., Chen, W., Acuna, D., Kautz, J., Fidler, S.: Neural light field estimation for street scenes with differentiable virtual object insertion. In: ECCV. pp. 380–397. Springer (2022)
33. Wang, Z., Shen, T., Gao, J., Huang, S., Munkberg, J., Hasselgren, J., Gojcic, Z., Chen, W., Fidler, S.: Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In: CVPR. pp. 8370–8380 (2023)
34. Wu, Z., Liu, T., Luo, L., Zhong, Z., Chen, J., Xiao, H., Hou, C., Lou, H., Chen, Y., Yang, R., et al.: Mars: An instance-aware, modular and realistic simulator for autonomous driving. In: CICAI (2023)
35. Xie, Z., Zhang, J., Li, W., Zhang, F., Zhang, L.: S-nerf: Neural radiance fields for street views. In: ICLR (2023)
36. Yan, X., Zou, Z., Feng, S., Zhu, H., Sun, H., Liu, H.X.: Learning naturalistic driving environment with statistical realism. Nature Communications **14**(1), 2037 (2023)
37. Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R.: Unisim: A neural closed-loop sensor simulator. In: CVPR. pp. 1389–1399 (2023)
38. Yongjie Zhu, Yinda Zhang, S.L., Shi, B.: Spatially-varying outdoor lighting estimation from intrinsics. In: CVPR (2021)
39. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4471–4480 (2019)
40. Yu, Y., Meka, A., Elgharib, M., Seidel, H.P., Theobalt, C., Smith, W.A.P.: Self-supervised outdoor scene relighting. In: Proc. of the European Conference on Computer Vision (ECCV) (2020)
41. Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: CVPR. pp. 5453–5462 (2021)