DreamCar: Leveraging Car-specific Prior for in-the-wild 3D Car Reconstruction

Xiaobiao Du, Haiyang Sun, Ming Lu, Tianqing Zhu, Xin Yu

***Abstract*—Self-driving industries usually employ professional artists to build exquisite 3D cars. However, it is expensive to craft large-scale digital assets. Since there are already numerous datasets available that contain a vast number of images of cars, we focus on reconstructing high-quality 3D car models from these datasets. However, these datasets only contain one side of cars in the forward-moving scene. We try to use the existing generative models to provide more supervision information, but they struggle to generalize well in cars since they are trained on synthetic datasets not car-specific. In addition, The reconstructed 3D car texture misaligns due to a large error in camera pose estimation when dealing with in-the-wild images. These restrictions make it challenging for previous methods to reconstruct complete 3D cars. To address these problems, we propose a novel method, named DreamCar, which can reconstruct high-quality 3D cars given a few images even a single image. To generalize the generative model, we collect a car dataset, named Car360, with over 5,600 vehicles. With this dataset, we make the generative model more robust to cars. We use this generative prior specific to the car to guide its reconstruction via Score Distillation Sampling. To further complement the supervision information, we utilize the geometric and appearance symmetry of cars. Finally, we propose a pose optimization method that rectifies poses to tackle texture misalignment. Extensive experiments demonstrate that our method significantly outperforms existing methods in reconstructing high-quality 3D cars. Our code is available.**

***Index Terms*—Self-Driving, 3D Reconstruction, 3D Generation**

I. Introduction

SELF-DRIVING vehicles are trained to safely interact with diverse environments and other vehicles. They must deal with infrequent dangerous situations, but collecting such dangerous data is challenging in the real-world setting [1]. Simulation offers a flexible solution for generating large-scale data safely. For the development of a reliable self-driving system, it is essential to have training data that encompasses as broad a spectrum of scenarios as possible. This necessitates a simulator equipped with a diverse and extensive collection of traffic assets, such as vehicles of varying sizes, shapes, and appearances, to ensure comprehensive coverage. Nevertheless, current self-driving simulators are limited as their 3D assets are created manually. The manual design of 3D assets is a labor-intensive and costly process, which limits scalability. Therefore, we intend to reconstruct large-scale high-quality 3D vehicle assets from existing self-driving datasets [2]–[4].

However, the process of producing such high-quality 3D cars from real-world sensor data faces 3 substantial challenges. (1) Captured Images that contain certain vehicles in self-driving datasets are often limited in number (ranging from one to five views). Particularly, in these images, only one side of the cars can be observed since they are captured in the moving forward scene. This scenario highlights a significant limitation: objects of interest are documented from restricted perspectives, with certain parts remaining unobserved. (2) Current large-scale 3D-aware diffusion models [5]–[7] trained on large-scale synthetic datasets, not car-specific, generalize poorly in cars, especially real-world cars, which hinders them from providing useful prior. (3) Since the ego vehicle is in the moving forward scene and may encounter vibration, the estimated poses in these self-driving datasets contain a quite large error, leading to texture misalignment while reconstructing 3D cars.

In this work, we propose a novel method for 3D car reconstruction in the moving forward scene, named DreamCar, which utilizes existing real sensor data collected by self-driving vehicles in the moving forward scenario to reconstruct a high-quality 3D asset library for realistic sensor simulation. Considering the challenging moving forward situation that only provides one to five supervision images, our method introduces more supervision information. In light of the mirror symmetry of the nature of cars, the image flip and pose symmetry techniques are leveraged in our method to generate mirror counterparts, increasing the amount of supervision reference images to two times.

To improve the generalization of the 3D-aware diffusion model to cars, we collect a high-quality car dataset, named Car360. As we can see in Figure 1 **b**, given input views, we use the current 3D-aware diffusion model, Zero-123-XL [5] to synthesize novel views. Initially, this model struggles to synthesize realistic car novel views. However, after training with our collected Car360 dataset, which contains over 5,600 synthetic cars with photorealistic textures, the performance of this model on car images significantly improved. This improvement emphasizes the importance of our datasets and underscores existing 3D-aware models that generalize poorly in cars.

To tackle the texture misalignment problem raised by camera pose error, we propose a pose optimization method. This method can be integrated into our method without explicit

This research is funded in part by ARC-Discovery grant (DP220100800 to XY) and ARC-DECRA grant (DE230100477 to XY). We thank all anonymous reviewers and editors for their constructive suggestions.

Xiaobiao Du is with the University of Technology Sydney, Australia (Email: xiaobiao.du@student.uts.edu.au)

Haiyang Sun and Ming Lu are with Li Auto Inc. and Intel Inc., China, respectively. (Email: sunhaiyang@lixiang.com, ming1.lu@intel.com)

Tianqing Zhu is currently a professor at city university of Macau, China. (Email: tianqing.zhu@uts.edu.au)

Xin Yu is with The University of Queensland. (Email: xin.yu@uq.edu.au)

Corresponding Author: Xin Yu

0000–0000/00$00.00 © 2021 IEEE

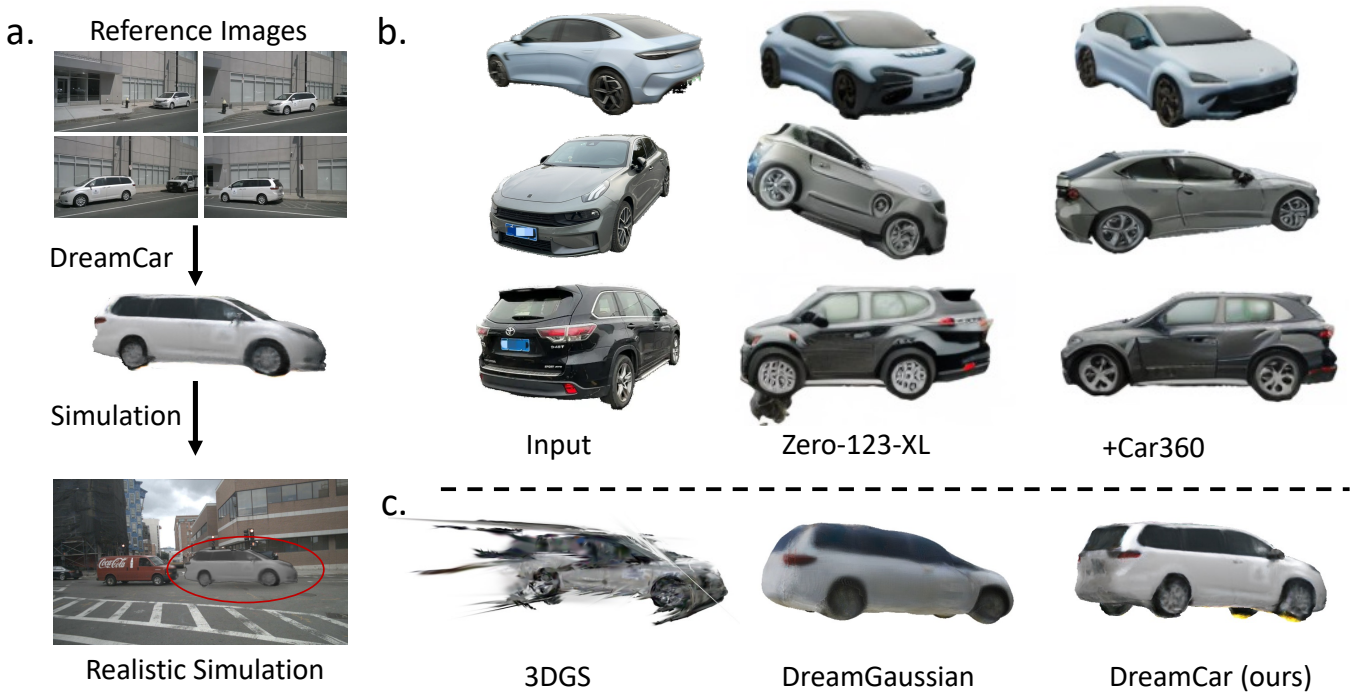


Fig. 1. **The illustration of the application scenario of our work and comparison with other methods. a.** Utilizing merely 4 reference images even less in the moving forward scene, we achieve the reconstruction of a complete 3D object, which is then simulated within a realistic scene. **b.** The novel view synthesis comparison provided by Zero-123-XL gradually trained on our Car360 dataset. **c.** The visual comparison of different 3D reconstruction methods.

supervision information. In particular, we design a Multilayer Perceptron (MLP), called PoseMLP, which takes the original pose and time-aware information as input to predict the offset for the correction to the original pose in the moving forward scenario.

As illustrated in Figure 1 **a**, we show the effect of our proposed method, DreamCar, within its application context. Despite the challenge posed by training with only four low-resolution and noisy reference images, with our proposed techniques, DreamCar is capable of accurately reconstructing a complete 3D object using 4 reference images, with both precise geometry and intricate texture. This method proves valuable for reconstructing large-scale 3D vehicle objects from existing autonomous driving datasets and enables realistic simulation across a variety of scenes. With our proposed techniques and dataset, Figure 1 **c** demonstrates that our method can reconstruct more realistic 3D cars and our collected dataset bridges the gap of our reconstructed cars to cars. To sum up, our contributions can be summarized as follows:

- We propose DreamCar, a novel 3D car reconstruction method, tailored for the moving forward scene. Our method integrates mirror symmetry, generative prior, and pose optimization techniques, which enable our method to reconstruct an intact 3D object in the self-driving scene even if it is only given a single reference image.
- To improve the generative model generalizing well in cars, we collect a Car dataset, termed Car360, with over 5,000 vehicles together, each adorned with photorealistic textures.
- To demonstrate our method, we evaluate our method in the large-scale self-driving dataset. We also evaluate it in our Car360 dataset. Extensive experiments demonstrate our method outperforms existing methods and is effective in reconstructing large-scale complete and high-quality 3D car objects.

II. Related Work

3D Reconstruction. The recent advancements [8]–[19] in 3D reconstruction and novel view synthesis significantly propel progress in the field. A particularly noteworthy approach is the Neural Radiance Fields (NeRF) [20] which represents the entire scene as a radiance field parameterized by an MLP. This method employs the volume rendering [21] to render the scene. However, the applicability of NeRF is constrained to specific contexts [22], such as synthetic environments, and it is limited to scenes within its bounds, struggling with out-of-bound scenarios. Several methods are proposed, including Mip-NeRF [23], Mip-NeRF 360 [22], TensoRF [24], and Instant-NGP [25], all aimed at extending the utility of NeRF to more general, in-the-wild scenes. The introduction of 3DGS [26] marks a further enhancement of the performance in 3D reconstruction, both in terms of quality and speed. Nonetheless, these methods encounter challenges in self-driving contexts, especially when tasked with reconstructing specific objects from a limited number of supervision images.

Diffusion Models for 3D Generation. Recent advancements in 2D diffusion models [27], [28], as highlighted by notable works such as Stable Diffusion, open new avenues for the generation of 3D objects, marking a significant leap forward in the field. DreamFusion [29] and SJC [30] pave the way by leveraging the capabilities of 2D text-to-image generation models to facilitate the creation of 3D shapes. This innovative approach inspires a wave of subsequent research, including

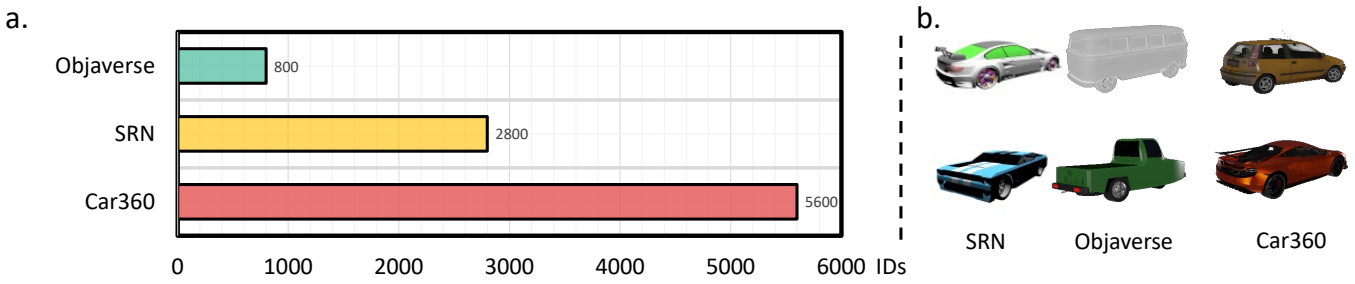


Fig. 2. **The illustration of different car datasets. a.** The comparison of the total number of vehicles across different datasets, highlights that our Car360 collection has the highest count, with 5,600 vehicles. **b.** The visual comparison of different car datasets. **c.** The illustration of the distribution of various vehicle categories, lighting conditions, and the number of captured views in our collected Car360 dataset.

text-to-3D methods [31]–[39] and single-image-to-3D methods [5], [40]–[46]. To deal with the multi-head problem, there are several methods [47]–[53] that are proposed to exploit multi-view information. To tackle the generated results with unrealistic texture, more powerful 3D generation methods [54]–[58] and improved diffusion strategies [46], [59]–[66] are proposed. To speed up the generation, some methods [45], [67]–[71] exploit Gaussian Splatting [26] as the main 3D model. Specifically, Dreamgaussian [72], as the representative Gaussian Splatting generating method, generates a 3D object in 40 seconds. On the contrary, for better generation, DreamCraft3D [73] adopts multi-stage training to achieve state-of-the-art generation results. Nevertheless, the application of these methods to realistic scenarios presents challenges, primarily because they heavily depend on generative models. This reliance often results in generated outcomes that lack plausibility when compared to realistic references.

III. Proposed Car360 Datasets

This work aims to reconstruct a complete 3D model from a limited number of images, typically ranging from one to five. However, relying solely on this supervision information is insufficient. Therefore, we integrate a generative prior from the recent large-scale 3D-aware diffusion model, Zero-123-XL [5] in our method. We found this model fails to generalize well in the realistic car subject as shown in Figure 1 **b**, attributed to its training on large-scale synthetic datasets, like Objaverse [74], [75], not car-specific. In this work, we collect a car dataset, named Car360, which contains 5,600 synthetic cars to enable our model to focus on cars and boost our model robust to realistic cars. Figure 2 **b** shows the reality comparison among different datasets.

As shown in Figure 2 **a** and **b**, we list the number of vehicles from the existing car dataset with 360-degree views and show some samples from them. We find that a large number of samples from SRN [76] and Objaverse [74], [75] are not realistic. Therefore, we collect a synthetic car dataset from Sketchfab [1], consisting of 2,000 3D vehicles, named Car360. Binding them with SRN and Objeaverse forms our Car360 dataset. This dataset contains more realistic car 3D models than other synthetic datasets. The generative model would be trained on this dataset to improve its robustness to cars.

[1]Sketchfab: https://sketchfab.com

IV. Method

A. Preliminaries

1) Score Distillation Sampling.: Score Distillation Sampling (SDS), as described in DreamFusion [77], is a representative method to optimize 3D representation via a pre-trained 2D diffusion model ϕ, like stable-diffusion [27]. In DreamFusion, MipNeRF [78] is adopted as the 3D representation with parameters θ subject to optimization. With g denoting the rendering function, the image produced, $I_r = g(\theta)$, emerges from this process. To make the rendered image I_r look like the sample generated from the diffusion model ϕ, SDS is proposed to leverage the generated prior from the diffusion model. SDS exploits the diffusion model to predict the sampled noise $\hat{\epsilon}_\phi$ given the noisy image z_t^r, text embedding y, and time step t, as an estimated score $\hat{\epsilon}_\phi(z_t^r; y, t)$. The method assesses the deviation between the rendered image I_r added with Gaussian noise ϵ to the and another noise $\hat{\epsilon}_\phi$ predicted by the diffusion model. This measurement guides the adjustment of the parameters θ. The process of calculating the gradient for this adjustment is described as follows:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}[w(t)\,(\hat{\epsilon}_\phi(z_t^r; y, t) - \epsilon)\,\frac{\partial I_r}{\partial \theta}], \quad (1)$$

where $w(t)$ is a weighting function. However, this method using the prompt with a certain view instruction to drive the diffusion model to generate a fixed view is rigid and only generates coarse views. To address this concern, Zero-123 [5] is proposed to exploit a conditioning image I_{cond} with a relative 3 Degrees of Freedom (DoF) pose Δp to generate a novel view. The integration of Zero-123 into SDS can be formulated as follows:

$$\nabla_\theta \mathcal{L}_{3D-SDS}(\phi, g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}[w(t)\,(\hat{\epsilon}_\phi(z_t^r; I_{cond}, \Delta p, t) - \epsilon)\,\frac{\partial I_r}{\partial \theta}]. \quad (2)$$

B. Camera Pose Processing

SfM [79] is a common method to reconstruct point clouds and estimate the camera pose according to a series of images in the real world. However, cameras on the ego vehicle only capture a few images of a certain object, making SfM fail to estimate the camera pose. Thus, we rely on the large-scale self-driving Object Detection dataset, Nuscenes [2] with clean

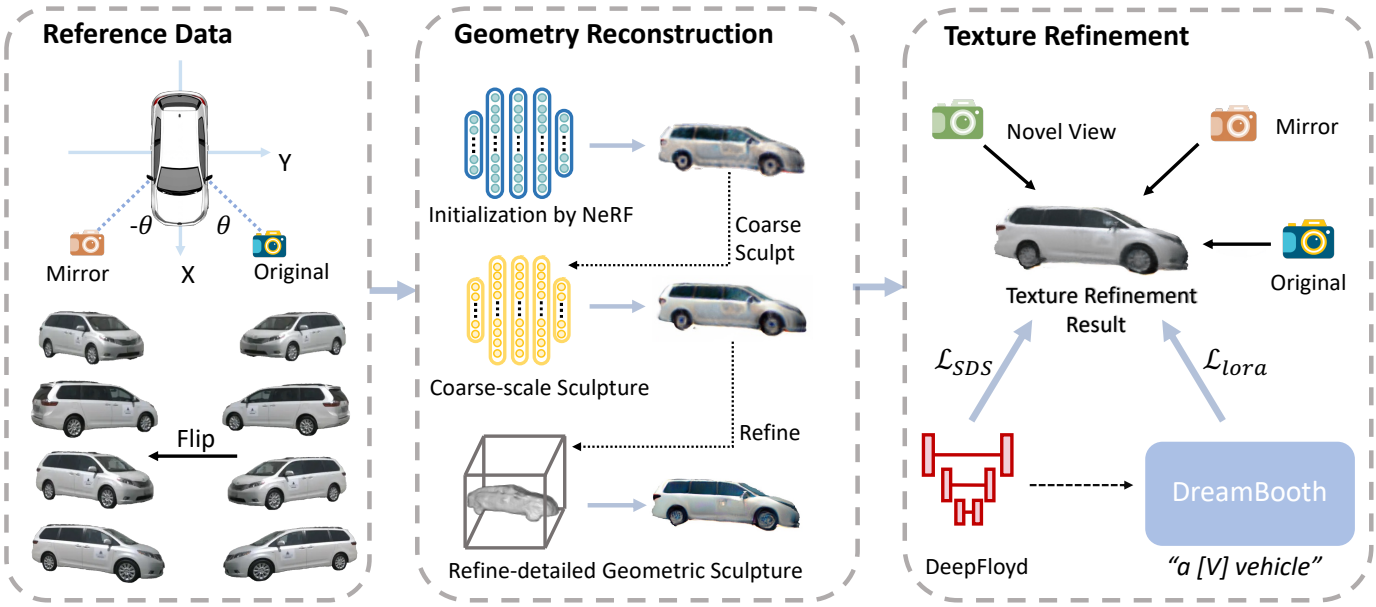


Fig. 3. **The illustration of our DreamCar.** Our method can be divided into geometry reconstruction and texture refinement stages. We input reference views (original reference views and their mirror counterparts) with generative prior guiding our 3D model to reconstruct 3D cars in all stages. In the geometry reconstruction stage, our method progressively sculpts fine geometry with coarse texture. In the texture refinement stage, we focus on the refinement of the appearance of the car with the DreamBooth technique.

annotations to extract the pose of captured vehicles. Given the camera poses p_{cam} and ego poses p_{ego} from the camera in the dataset like Nuscenes [2], these poses can all be described by Lie group $SE(3)$ [80]. The Lie group can be formulated as follows:

$$SE(3) = \left\{ \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} | \mathbf{R} \in SO(3), \mathbf{T} \in \mathbb{R}^3 \right\}, \quad (3)$$

where $\mathbf{R}$ and $\mathbf{T}$ denote the rotation matrix and translation. With these two poses, we can easily get the camera-to-world pose $p_{c2w} = p_{ego}p_{cam}$. Since we target to reconstruct an object, not the whole scene, so we need to transfer the current reference system to "object-centric". We extract the pose of a certain object p_{obj} by using the bounding box annotation. The central point of the 3D bounding box and its orientation can represent the translation and rotation matrix as the pose of an object, as p_{obj}. Therefore, we transfer the camera-to-world pose to "object-centric" $p_{c2obj} = p_{obj}^{-1}p_{c2w}$. The inverse of the pose can be calculated by:

$$SE(3)^{-1} = \left\{ \begin{bmatrix} \mathbf{R}^T & -\mathbf{R}^T\mathbf{T} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4\times4} | \mathbf{R} \in SO(3), \mathbf{T} \in \mathbb{R}^3 \right\}. \quad (4)$$

Since we aim at reconstructing a 3D object in large-scale self-driving datasets, the number of images of a certain object is about one to five, making existing methods fail to reconstruct a high-quality 3D object. Thus, we reconstruct a 3D object by means of the generated prior from SDS. SDS replies on the large-scale diffusion models [5], [27] to provide generated information. However, these generated models all assume the object is in the center of an image. Therefore, we recenter the input images and adjust the camera pose to look at the original point. The final pose p would be obtained by adjusting the rotation matrix of pose p_{c2obj} to look at the coordinate original point.

C. Pre-processing and Attainment of Mirror Symmetry

Given a few images $I = \{I_1, I_2, \cdots I_n\}$ each capturing a certain vehicle from an autonomous driving dataset with the number of images $|I|$ ranging from one to five, our objective is to reconstruct a 3D object characterized by precise geometry and clean texture. Remarkably, our method is capable of reconstructing a high-quality 3D model from a few images even a single image, despite the challenging constraints posed by such few supervision data. Our initial step involves employing the Segment Anything Model (SAM) [81] to segment the vehicle from the background, focusing our reconstruction solely on the vehicle rather than the entire scene. To address the scarcity of supervision information, we expand this limited training data by exploiting mirror symmetry to enrich the information base for the reconstruction process.

There are plenty of objects that have symmetry in nature, and vehicles are no exception. Generally speaking, the left and right side of a car has the same appearance and keep symmetry. Considering this significant feature, we first expand the number of reference images by the mirror flip. Denoting the image flip function as F_{flip}, the mirror symmetry image I_i' can be obtained by simply flipping the image $I_i' = F_{flip}(I_i)$. The leftmost part of Fig. 3 demonstrates how we obtain the image and pose mirror. To obtain the mirror camera pose, taking the camera pose of the Colmap format [79] for instance, the car orients along the x-axis, so we just need to negative the y component of the translation and mirror the rotation matrix. With the obtained mirror images and poses, the reference images can be expanded to two times $I = \{I_1, I_2, \cdots I_n, I_1', I_2', \cdots I_n'\}$.

D. Geometry Reconstruction

As shown in Fig. 3, our method can be divided into the geometry reconstruction and texture refinement stages. In the geometry reconstruction stage, we focus on sculpting a 3D object with intact and precise 3D geometry and coarse texture such that it matches reference images I at reference views and maintains plausibility at various views.

To reach this target, we choose to progressively sculpt fine structural geometry by three different 3D models. NeRF is first adopted to reconstruct a coarse geometry and shape. Then, we use the result of NeRF as initialization for the Neus [82]. Once the coarse-scale sculpture by Neus is completed, we use DMTET [83] initialized by Neus to reconstruct a fine geometry. To leverage most of the absolute supervision information provided by the reference image, we penalize for the foreground discrepancy observed between the rendered image and the reference and their mask through

$$\mathcal{L}_{\text{rgb}} = \left\| m_i \odot (I_i - g(\theta; p_i)) \right\|_2, \quad \mathcal{L}_{\text{mask}} = \left\| m_i - g_m(\theta; p_i) \right\|_2, \tag{5}$$

at the i-th reference pose p_i. We only compute the loss within the foreground region through the reference mask m_i. At the same time, we also compute the mask loss to encourage the outline convergence, where g_m renders the silhouette. In addition, drawing inspiration from [84], we compute both depth and normal losses to thoroughly utilize the geometric prior derived from the reference image, thereby guaranteeing geometry matching with the reference. The depth and normal loss can be computed as:

$$\mathcal{L}_{\text{depth}} = -\frac{\text{cov}(d_i, g_d(\theta; p_i))}{\sigma(d_i)\sigma(g_d(\theta; p_i))}, \quad \mathcal{L}_{\text{normal}} = -\frac{n_i \cdot g_n(\theta; p_i)}{\|n_i\|_2 \cdot \|g_n(\theta; p_i)\|_2}, \tag{6}$$

where $\text{cov}(\cdot)$ and $\sigma(\cdot)$ denote the covariance and variance operators respectively. We obtain the reference depth d_i and the normal n_i by leveraging the off-the-shelf network [85]. The rendered depth and normal are obtained by $g_d(\theta; p_i)$ and $g_n(\theta; p_i)$. To the depth loss, we adopt the negative Pearson correlation $\mathcal{L}_{\text{depth}}$ to address discrepancies in the depth scale. To normal loss, we simply compute its cosine similarity to ensure the fitting of the rendered normal and the reference one. However, this supervision information is not enough for the reconstruction of a complete 3D object. Thus, we employ the generative prior to enhance the reconstruction of distinct views given references.

Generative Prior. The recent 3D-aware generative model, Zero-123 [5], has been trained on a vast array of 3D objects, endowing it with an innate understanding of 3D poses up to three degrees of freedom (3DoF). In this work, we leverage our collected Car360 dataset to enhance the generalization of the model to real-world cars. This model is then employed to generate detailed information that aids in the reconstruction of a comprehensive 3D model. Moreover, we incorporate a 2D generative model to augment the generative capabilities, given its training on a broader and more diverse dataset compared to those available to 3D-aware generative models. Consequently, we employ a mixed SDS loss that mixes the generative priors from both 2D and 3D perspectives. This mixed SDS loss is utilized to distill the gradient as follows:

$$\nabla_\theta \mathcal{L}_{mix}(\phi, g(\theta)) = \nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, g(\theta)) + \nabla_\theta \mathcal{L}_{\text{3D-SDS}}(\phi, g(\theta)). \tag{7}$$

When computing $\mathcal{L}_{\text{SDS}}$, the DeepFloyd IF base model [86] [27] is adopted as 2D prior to capture coarse geometry. Therefore, the total geometry reconstruction loss can be formulated as:

$$\begin{aligned} \mathcal{L}_{geo} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{depth}\mathcal{L}_{depth} \\ + \lambda_{normal}\mathcal{L}_{normal} + \lambda_{mix}\mathcal{L}_{mix}, \end{aligned} \tag{8}$$

where $\mathcal{L}_{rgb}, \mathcal{L}_{mask}, \mathcal{L}_{depth}, \mathcal{L}_{normal}$ are reference losses, and $\mathcal{L}_{mix}$ means the guidance loss introducing generative prior. All λ are weights for all losses.

E. Texture Refinement

The initial stage of geometry reconstruction yields precise geometry, yet the resulting textures are overly smooth and lack sharpness. This issue stems from the reliance on our 2D generative prior model, which generates images at a low resolution and exhibits considerable inconsistency when producing the specified subject. In this stage, we would focus on the refinement of texture.

To enhance the realism of the textures, we integrate the Stable Diffusion model [27] at this stage to provide gradients of higher resolution. Additionally, during this learning phase, we continue to utilize the aforementioned reference loss, given its value as an effective source of supervision information. To improve the stability of the generative model, we utilize the recently proposed DreamBooth [87] and Lora [51], [88] techniques. Specifically, we set the text prompts containing the class of vehicles and a unique identifier (e.g., "a [V] vehicle" in Fig. 3) to enable the generation of the diffusion model stable to an identity. To the problem of blurry and over-statute texture, we introduce Lora, a large-scale model fine-tuning technique, which significantly enhances texture realism during the gradient distillation phase. Consequently, at this stage, we incorporate two additional loss functions, $L_{dreambooth}$ and L_{lora}, to achieve a more realistic texture generation.

Pose Optimization. In the moving forward scene, the estimated pose to a certain captured object exists a quite large error since it is in a dynamic environment and may encounter vibration. This problem would lead to texture misalignment in the process of the reconstruction of a fine 3D object. To tackle this problem, we propose a pose optimization method, which accepts the time frame and the original pose as input to predict the offset for the correction to the original pose. Specifically, we design an MLP, called PoseMLP $N_{posemlp}$, with 3 layers of 256 hidden neural units. To the rotation matrix, we do not simply input the whole one but rather quantify the rotation of each axis as input. Therefore, the total input for this network is 6D vectors and the time frame. Our pose optimization can be formulated as:

$$\hat{p_i} = N_{posemlp}(i, p_i) + p_i, \tag{9}$$

where p_i denotes the i-th pose. With this time-related information as input, this network is endowed with time-aware ability

TABLE I
THE COMPARISON OF QUANTITATIVE 3D RECONSTRUCTION METRICS IN THE TEST SET FROM THE CAR360 DATASET. "STANDARD" MEANS THE EVALUATION IN STANDARD TESTING VIEWS. "MIRROR" DENOTES THE EVALUATION OF THE MIRROR VERSION OF THE TESTING VIEW TO MEASURE THE COMPLETENESS OF RECONSTRUCTED 3D CARS. INFER. SPEED (N/S) MEANS THE NUMBER OF RENDERED IMAGES PER SECOND. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Testing View	Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Standard	TensoRF [24]	0.1224	9.23	0.3685	0.5932	0.91
	Instant-NGP [25]	0.1127	9.48	0.3837	0.5491	0.87
	3DGS [26]	0.0933	10.06	0.4027	0.5022	0.67
	DreamCar (Ours)	**0.0297**	**15.44**	**0.6894**	**0.2281**	**0.23**
Mirror	TensoRF [24]	0.2367	6.27	0.3092	0.6121	1.92
	Instant-NGP [25]	0.1929	7.21	0.3281	0.6075	1.59
	3DGS [26]	0.1467	7.83	0.3331	0.5904	1.31
	DreamCar (Ours)	**0.0429**	**14.71**	**0.6761**	**0.2591**	**0.73**

TABLE II
THE COMPARISON OF POINT CLOUDS AND DETECTION METRICS IN THE TEST SET FROM THE CAR360 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Geometry	3DGS [26]	Zero-123 [5]	DreamGaussian [72]	DreamCraft3D [73]	**Dreamcar**
L_2-error ↓	0.24	0.19	0.15	0.14	**0.13**
Hit rate ↑	66.38%	86.19%	89.21%	91.37%	**93.48%**
Chamfer ↓	9.66	2.14	0.87	0.41	**0.24**
Hausdorff ↓	1.8418	1.28	1.45	1.39	**0.98**
map@0.5 ↑	0.04	0.23	0.31	0.67	**0.83**

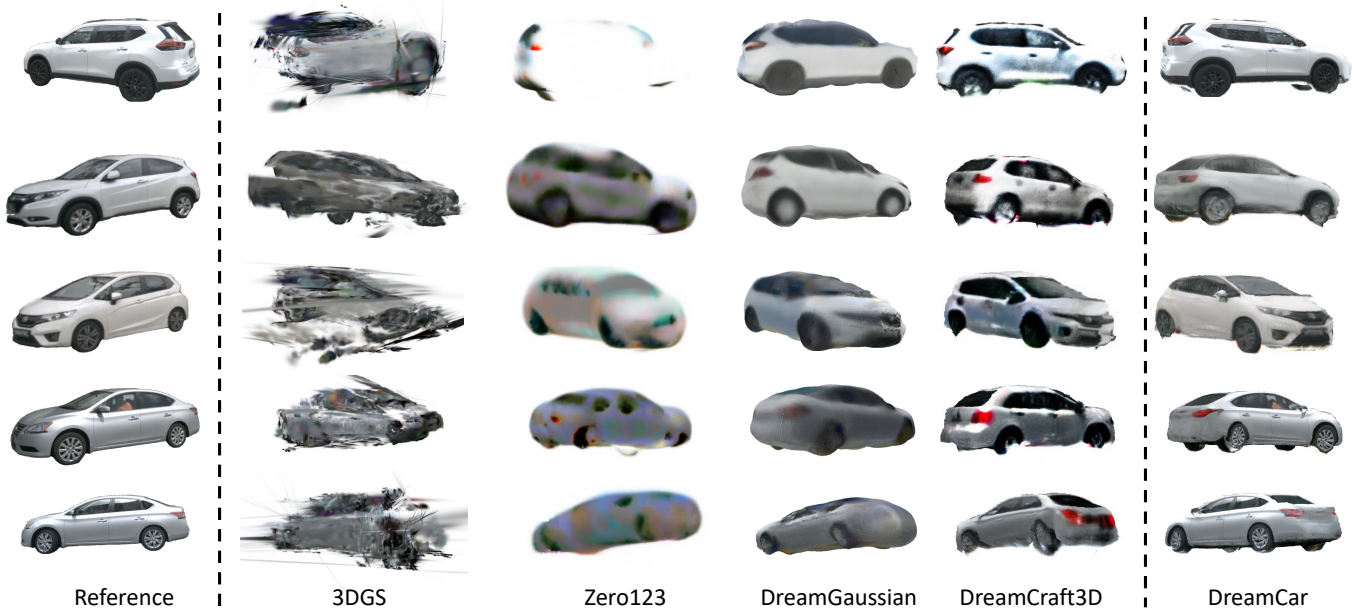


Fig. 4. **Qualitative evaluation of 3D reconstruction on the Nuscenes Dataset [2].** The renderings are provided from various viewpoints distinct from the reference image, illustrating the completeness of the reconstructed 3D models.

in the moving forward scene. Note that our proposed PoseMLP can be integrated into our method and does not need explicit supervision information. Therefore, the texture refinement loss can be represented as

$$\mathcal{L}_{tex} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{normal}\mathcal{L}_{normal} + \lambda_{3d}\mathcal{L}_{3D-SDS} + \lambda_{booth}\mathcal{L}_{dreambooth} + \lambda_{lora}\mathcal{L}_{lora}, \tag{10}$$

where λ_{rgb} would be the largest one since it provides absolution appearance supervision for better texture refinement.

V. EXPERIMENTS

A. Implementations

During the geometry reconstruction stage, we opt for a low resolution of 128 for both NeRF and Neus, aiming to expedite training time while achieving a coarse geometric structure. In the subsequent stages of geometry and texture refinement, we increase the resolution to 1024. This adjustment is designed to capture more intricate details in both geometry and texture. Furthermore, we scale the radius of the high-quality pose to 2 and modify the field-of-view (FOV) angles to be consistent with the methodology employed by Zero-123-XL [5]. Regard-

TABLE III
THE COMPARISON OF QUANTITATIVE 3D RECONSTRUCTION METRICS IN 100 VEHICLES FROM THE NUSCENES DATASET. "STANDARD" MEANS WE RANDOMLY CHOOSE ONE IMAGE OF EACH VEHICLE AS THE TEST IMAGE. "MIRROR" DENOTES THE MIRROR VERSION OF THE TEST IMAGE, EMPLOYED TO VERIFY THE COMPLETENESS OF THE 3D MODEL. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Testing View	Method	MSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Standard	NeRF [20]	0.1341	8.74	0.3557	0.6019	1.31
	TensoRF [24]	0.1224	9.23	0.3685	0.5932	0.91
	Instant-NGP [25]	0.1127	9.48	0.3837	0.5491	0.87
	Zip-NeRF [89]	0.1093	9.72	0.3861	0.5134	0.52
	3DGS [26]	0.0933	10.06	0.4027	0.5022	0.67
	DreamCar (Ours)	**0.0297**	**15.44**	**0.6894**	**0.2281**	**0.23**
Mirror	NeRf [20]	0.2481	5.13	0.2714	0.6934	2.36
	TensoRF [24]	0.2367	6.27	0.3092	0.6121	1.92
	Instant-NGP [25]	0.1929	7.21	0.3281	0.6075	1.59
	Zip-NeRF [89]	0.1541	7.42	0.3357	0.6012	1.45
	3DGS [26]	0.1467	7.83	0.3331	0.5904	1.31
	DreamCar (Ours)	**0.0429**	**14.71**	**0.6761**	**0.2591**	**0.73**

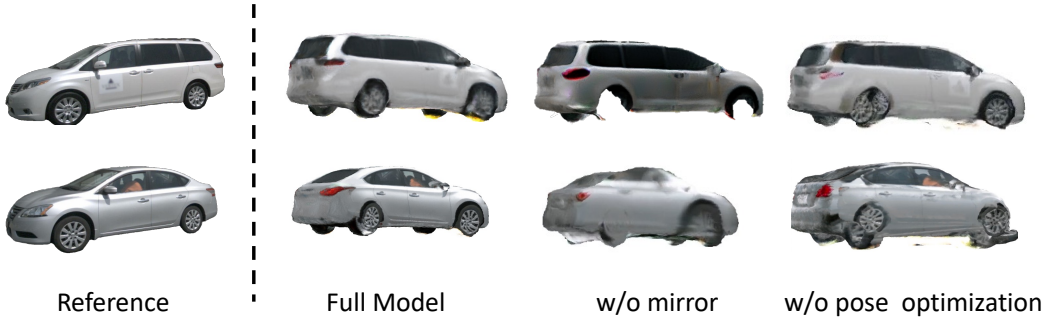


Fig. 5. **The ablation study of our proposed method.** Given the leftmost references, we ablate the mirror and pose optimization techniques to demonstrate our method.

ing the generative model, we finetune Zero-123-XL [5] on our collected Car360 dataset to improve the model generating more reliable car-specific prior.

In the geometry reconstruction stage, we set $\mathcal{L}_{rgb} = 1000$, $\mathcal{L}_{mask} = 100$, $\mathcal{L}_{depth} = 0.05$, $\mathcal{L}_{normal} = 1$, $\mathcal{L}_{mix} = 0.1$. In the texture refinement stage, to maximize the utilization of reference data, we set $\mathcal{L}_{rgb} = 10000$, $\lambda_{3d} = 0.1$, $\lambda_{booth} = 0.1$, $\lambda_{lora} = 0.01$. The pose optimization is only applied to the NeRF reconstruction stage with a learning rate of 10^{-5}. Note that our pose optimization is integrated into NeRF training without explicit pose supervision. Each stage is trained in 5000 epochs. Completing all stages for a vehicle requires approximately 2 hours on an RTX 3090Ti GPU Card.

Camera Processing. Our method leverages Score Distillation Sampling and thus it requires the object centrally located within the conditioning image. Therefore, we re-center the object in an image based on its bounding box.

B. Qualitative and Quantitative Results

Datasets. Our target is to reconstruct complete and fine 3D cars from the real-world scene. Therefore, we select 100 instances from our collected Car360 dataset for evaluation. To imitate the moving forward scene, we randomly select 3 sparse views and 1 view from one side of each vehicle as training and testing views, respectively. To further demonstrate our method, we also extract 100 vehicles from the Nuscenes [2] dataset for further evaluation. The number of captured images for each vehicle ranges from 2 to 5. This dataset is documented in a moving-forward scenario, matching the realistic scene.

Evaluation Settings and Metrics. We evaluate the reconstruction quality by using image-based and lidar-based metrics. We use the common image-based metrics for rendered views, such as PSNR, SSIM, LPIPS, and FID. In image-based metrics, we denote the "Standard" as the evaluation in testing views. We also evaluate the mirror testing views denoted as "Mirror" in Table I to compare which method can reconstruct more complete 3D cars. For lidar-based metrics, we adopt per-ray L_2 error, Hit rate, Chamfer distance, and Hausdorff to measure the completeness of reconstructed 3D cars. Note that lidars in Nuscenes only scan one side of cars since it is in the moving forward scenes, so we would not evaluate this incomplete point cloud in Nuscenes. On the contrary, our Car360 contains a complete point cloud for each vehicle, which would be used to evaluate the lidar-based metrics. In addition to these metrics, we also adopt map@0.5 as a detection metric to evaluate if existing detectors can detect reconstructed cars. Higher map@0.5 means the reconstructed results are more photorealistic.

Car360. As indicated in Table I, we adopt the recent 3D reconstruction methods, such as TensoRF [24], Instant-NGP [25], and 3DGS [26] to compare with our proposed DreamCar. Notably, 3DGS [26], despite being the recent state-of-the-art method, shows extremely bad quantitative metrics than ours.

Fig. 6. **The comparison of novel view synthesis across 3D-aware diffusion models.** The red circle highlights the unreasonable generation of Stable-Zero123 [7].

Fig. 7. **Car simulation in the original Nuscenes scenes.** The red circle indicates the simulated car.

When testing on the mirror views, their results are worse. As shown in Table II, our proposed method obtains the best results against other methods in lidar-based metrics. These methods struggle to fully reconstruct 3D objects when the test view significantly deviates from the reference. We also compare our method with the recent 3D generation methods, such as Zero-123 [5], DreamGaussian [72], and DreamCraft3D [73]. For a fair comparison, we modify these models to accept multi-view inputs. Specifically, Dreamfusion [29] is utilized for Zero-123 in the 3D reconstruction process. We also use existing detectors YOLOv8-l [90] to evaluate which method can reconstruct more photorealistic cars to be detected. Specifically, we copy and paste the rendered images into the backgrounds of Nuscenes to form testing sets. Then, we use YOLOv8-l to detect these testing sets and only evaluate on the rendered objects. The above comparisons indicate our DreamCar can produce better results than previous methods.

Nuscenes. Table III depicts the quantitative results, where "Standard" and "Mirror" rows all show our method is superior to other 3D reconstruction methods. Previous 3D reconstruction methods fail to reconstruct cars given such a few images. Our results in Nuscenes are worse than ours in Car360 since the images in Nuscenes are low-resolution, blurry, and noisy, which indicates this dataset is more challenging. As shown in Figure 4, we show the visual comparison of 3D reconstruction with the recent 3D generation methods. Note that these methods are all modified to accept multi-view inputs. 3DGS can not reconstruct a complete 3D car given a few supervision images. Zero-123 struggles to reconstruct a complete 3D geometry and detailed texture. The geometry reconstructed by DreamGussian is like an ellipsoid. Moreover, its texture is very blurry and over-smoothing. DreamCraft3D, although a recent state-of-the-art 3D generation model, sometimes misses details such as vehicle wheels and produces textures that are overly saturated and lack photorealism. Since our proposed method exploits mirror symmetry and has more powerful generalization in real-world cars, our DreamCar can reconstruct real-world cars from the moving forward scene. These comparisons demonstrate the applicability of our method in realistic scenes for the production of large-scale 3D car assets.

C. Ablation Study

In Figure 5, we show the ablation study through ablating our proposed mirror and pose optimization, respectively. When

our full model does not use mirror symmetry, the generated geometry is incomplete and the texture looks different with the reference. Without pose optimization, the generated texture displays misalignments in the vehicle wheels. To highlight the benefits of our Car360 dataset, we conduct a novel view synthesis comparison in Figure 6, with other state-of-the-art 3D-aware diffusion models, such as Zero-123-XL [5], SyncDreamer [6], and Stable-Zero123 [7]. As we can see, Zero-123-XL and SyncDreamer generate unrealistic results and seem to pattern collapse. Stable-Zero123 also generates unreasonable features, as the vehicle window is not matched with the real one. Moreover, it seems the generating pose has a few deviations. The generative model trained on our Car360 dataset obtains powerful generalization to real cars and accomplishes better visual results than others. These comparisons demonstrate the necessity of our collected Car360 dataset.

D. Car Simulation.

As illustrated in Figure 7, we present the visualization of the car simulation. We first use our method to reconstruct a bulk of 3D car models. Then, the reconstructed 3D car models are employed for simulation within Nuscenes [2] scenes. Utilizing these 3D cars allows for the simulation of large-scale, highly hazardous scenarios in a cost-effective way.

VI. Conclusion

In this work, we propose a groundbreaking 3D car reconstruction method, named DreamCar. The success of DreamCar can be attributed to its innovative application of mirror symmetry, car-specific generative prior, and pose optimization techniques. Specifically, the usage of the collected Car360 datasets improves the generalization of our method to real-world cars. We demonstrate our DreamCar can effectively reconstruct large-scale exquisite 3D car assets and outperform other state-of-the-art methods.

References

[1] S. Tan, K. Wong, S. Wang, S. Manivasagam, M. Ren, and R. Urtasun, "Scenegen: Learning to generate realistic traffic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 892–901, 2021.

[2] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

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

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9298–9309, 2023.

[6] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," *arXiv preprint arXiv:2309.03453*, 2023.

[7] S. AI, "Stable Zero123: Quality 3d object generation from single images." https://stability.ai/news/stable-zero123-3d-generation, 2023.

[8] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," *arXiv preprint arXiv:2010.07492*, 2020.

[9] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8248–8258, 2022.

[10] W. Xian, J.-B. Huang, J. Kopf, and C. Kim, "Space-time neural irradiance fields for free-viewpoint video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9421–9431, 2021.

[11] M. Boss, R. Braun, V. Jampani, J. T. Barron, C. Liu, and H. Lensch, "Nerd: Neural reflectance decomposition from image collections," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12684–12694, 2021.

[12] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5610–5619, 2021.

[13] C. Reiser, R. Szeliski, D. Verbin, P. Srinivasan, B. Mildenhall, A. Geiger, J. Barron, and P. Hedman, "Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–12, 2023.

[14] Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "Nerf-editing: geometry editing of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18353–18364, 2022.

[15] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14346–14355, 2021.

[16] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

[17] Z. Chen, T. Funkhouser, P. Hedman, and A. Tagliasacchi, "Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16569–16578, 2023.

[18] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5741–5751, 2021.

[19] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, *et al.*, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–12, 2023.

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

[21] J. T. Kajiya and B. P. Von Herzen, "Ray tracing volume densities," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 165–174, 1984.

[22] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5470–5479, 2022.

[23] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.

[24] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," *arXiv preprint arXiv:2203.09517*, 2022.

[25] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *arXiv:2201.05989*, Jan. 2022.

[26] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (ToG)*, vol. 42, no. 4, pp. 1–14, 2023.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[28] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[29] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[30] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich, "Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12619–12629, 2023.

[31] Z. Chen, F. Wang, and H. Liu, "Text-to-3d using gaussian splatting," *arXiv preprint arXiv:2309.16585*, 2023.

[32] H. Seo, H. Kim, G. Kim, and S. Y. Chun, "Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model," *arXiv preprint arXiv:2304.02827*, 2023.

[33] C. Yu, Q. Zhou, J. Li, Z. Zhang, Z. Wang, and F. Wang, "Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation," *arXiv preprint arXiv:2307.13908*, 2023.

[34] J. Seo, W. Jang, M.-S. Kwak, J. Ko, H. Kim, J. Kim, J.-H. Kim, J. Lee, and S. Kim, "Let 2d diffusion model know 3d-consistency for robust text-to-3d generation," *arXiv preprint arXiv:2303.07937*, 2023.

[35] C. Tsalicoglou, F. Manhardt, A. Tonioni, M. Niemeyer, and F. Tombari, "Textmesh: Generation of realistic 3d meshes from text prompts," *arXiv preprint arXiv:2304.12439*, 2023.

[36] M. Armandpour, H. Zheng, A. Sadeghian, A. Sadeghian, and M. Zhou, "Re-imagine the negative prompt algorithm: Transform 2d diffusion into 3d, alleviate janus problem and beyond," *arXiv preprint arXiv:2304.04968*, 2023.

[37] Y. Chen, C. Zhang, X. Yang, Z. Cai, G. Yu, L. Yang, and G. Lin, "It3d: Improved text-to-3d generation with explicit view synthesis," *arXiv preprint arXiv:2308.11473*, 2023.

[38] D. Xu, Y. Jiang, P. Wang, Z. Fan, Y. Wang, and Z. Wang, "Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 views," *arXiv e-prints*, pp. arXiv–2211, 2022.

[39] X. Cheng, T. Yang, J. Wang, Y. Li, L. Zhang, J. Zhang, and L. Yuan, "Progressive3d: Progressively local editing for text-to-3d content creation with complex semantic prompts," *arXiv preprint arXiv:2310.11784*, 2023.

[40] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, *et al.*, "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors," *arXiv preprint arXiv:2306.17843*, 2023.

[41] W. Yu, L. Yuan, Y.-P. Cao, X. Gao, X. Li, L. Quan, Y. Shan, and Y. Tian, "Hifi-123: Towards high-fidelity one image to 3d content generation," *arXiv preprint arXiv:2310.06744*, 2023.

[42] Q. Shen, X. Yang, and X. Wang, "Anything-3d: Towards single-view anything reconstruction in the wild," *arXiv preprint arXiv:2304.10261*, 2023.

[43] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion: 360deg reconstruction of any object from a single image," in *CVPR*, 2023.

[44] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen, "Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior," in *ICCV*, 2023.

[45] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. Kreis, "Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models," *arXiv preprint arXiv:2312.13763*, 2023.

[46] Y. Ma, Y. Fan, J. Ji, H. Wang, X. Sun, G. Jiang, A. Shu, and R. Ji, "X-dreamer: Creating high-quality 3d content by bridging the domain gap between text-to-2d and text-to-3d generation," *arXiv preprint arXiv:2312.00085*, 2023.

[47] Y. Shi, P. Wang, J. Ye, M. Long, K. Li, and X. Yang, "Mvdream: Multi-view diffusion for 3d generation," *arXiv preprint arXiv:2308.16512*, 2023.

[48] W. Li, R. Chen, X. Chen, and P. Tan, "Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3d," *arXiv preprint arXiv:2310.02596*, 2023.

[49] T. Huang, Y. Zeng, Z. Zhang, W. Xu, H. Xu, S. Xu, R. W. Lau, and W. Zuo, "Dreamcontrol: Control-based text-to-3d generation with 3d self-prior," *arXiv preprint arXiv:2312.06439*, 2023.

[50] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," *arXiv preprint arXiv:2310.15008*, 2023.

[51] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *arXiv preprint arXiv:2305.16213*, 2023.

[52] J. Zhang, Z. Tang, Y. Pang, X. Cheng, P. Jin, Y. Wei, W. Yu, M. Ning, and L. Yuan, "Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting," *arXiv preprint arXiv:2312.13271*, 2023.

[53] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data," *arXiv preprint arXiv:2306.07881*, 2023.

[54] M. Liu, R. Shi, L. Chen, Z. Zhang, C. Xu, X. Wei, H. Chen, C. Zeng, J. Gu, and H. Su, "One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion," *arXiv preprint arXiv:2311.07885*, 2023.

[55] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," 2023.

[56] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron, *et al.*, "Dreambooth3d: Subject-driven text-to-3d generation," *arXiv preprint arXiv:2303.13508*, 2023.

[57] R. Chen, Y. Chen, N. Jiao, and K. Jia, "Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation," *arXiv preprint arXiv:2303.13873*, 2023.

[58] M. Liu, C. Xu, H. Jin, L. Chen, M. Varma T, Z. Xu, and H. Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[59] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *CVPR*, 2023.

[60] Z. Liu, Y. Li, Y. Lin, X. Yu, S. Peng, Y.-P. Cao, X. Qi, X. Huang, D. Liang, and W. Ouyang, "Unidream: Unifying diffusion priors for relightable text-to-3d generation," *arXiv preprint arXiv:2312.08754*, 2023.

[61] M. Zhao, C. Zhao, X. Liang, L. Li, Z. Zhao, Z. Hu, C. Fan, and X. Yu, "Efficientdreamer: High-fidelity and robust 3d creation via orthogonal-view diffusion prior," *arXiv preprint arXiv:2308.13223*, 2023.

[62] J. Zhu and P. Zhuang, "Hifa: High-fidelity text-to-3d with advanced diffusion guidance," *arXiv preprint arXiv:2305.18766*, 2023.

[63] Y. Huang, J. Wang, Y. Shi, X. Qi, Z.-J. Zha, and L. Zhang, "Dreamtime: An improved optimization strategy for text-to-3d content creation," *arXiv preprint arXiv:2306.12422*, 2023.

[64] J. Wu, X. Gao, X. Liu, Z. Shen, C. Zhao, H. Feng, J. Liu, and E. Ding, "Hd-fusion: Detailed text-to-3d generation leveraging multiple noise estimation," *arXiv preprint arXiv:2307.16183*, 2023.

[65] Y. Jiang, H. Tang, J.-H. R. Chang, L. Song, Z. Wang, and L. Cao, "Efficient-3dim: Learning a generalizable single-image novel-view synthesizer in one day," *arXiv preprint arXiv:2310.03015*, 2023.

[66] Y. Chen, Z. Li, and P. Liu, "Et3d: Efficient text-to-3d generation via multi-view distillation," *arXiv preprint arXiv:2311.15561*, 2023.

[67] T. Yi, J. Fang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang, "Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors," *arXiv preprint arXiv:2310.08529*, 2023.

[68] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, and Z. Liu, "Humangaussian: Text-driven 3d human generation with gaussian splatting," *arXiv preprint arXiv:2311.17061*, 2023.

[69] J. Chung, S. Lee, H. Nam, J. Lee, and K. M. Lee, "Luciddreamer: Domain-free generation of 3d gaussian splatting scenes," *arXiv preprint arXiv:2311.13384*, 2023.

[70] D. Xu, Y. Yuan, M. Mardani, S. Liu, J. Song, Z. Wang, and A. Vahdat, "Agg: Amortized generative 3d gaussians for single image to 3d," *arXiv preprint arXiv:2401.04099*, 2024.

[71] S. Szymanowicz, C. Rupprecht, and A. Vedaldi, "Splatter image: Ultra-fast single-view 3d reconstruction," *arXiv preprint arXiv:2312.13150*, 2023.

[72] J. Tang, J. Ren, H. Zhou, Z. Liu, and G. Zeng, "Dreamgaussian: Generative gaussian splatting for efficient 3d content creation," *arXiv preprint arXiv:2309.16653*, 2023.

[73] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu, "Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior," *arXiv preprint arXiv:2310.16818*, 2023.

[74] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

[75] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

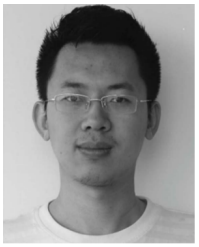
[76] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.
[77] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," *arXiv preprint arXiv:2209.14988*, 2022.
[78] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5855–5864, 2021.
[79] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.
[80] A. Mishra, S. P. Awate, A. Banerjee, N. El-Zehiry, S. Kurtek, S. J. McKenna, and A. Tannenbaum, "Lie groups in computer vision and image processing: A survey," *Journal of Mathematical Imaging and Vision*, vol. 47, no. 3, pp. 209–252, 2013.
[81] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
[82] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.
[83] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.
[84] C. Deng, C. Jiang, C. R. Qi, X. Yan, Y. Zhou, L. Guibas, D. Anguelov, *et al.*, "Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20637–20647, 2023.
[85] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10786–10796, 2021.
[86] A. Shonenkov, M. Konstantinov, D. Bakshandaeva, C. Schuhmann, K. Ivanova, and N. Klokova, "DeepFloyd IF: A modular cascaded diffusion model." https://github.com/deep-floyd/IF/tree/develop, 2023.
[87] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.
[88] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.
[89] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Zip-nerf: Anti-aliased grid-based neural radiance fields," *ICCV*, 2023.
[90] Ultralytics, "YOLOv8: A cutting-edge and state-of-the-art (sota) model that builds upon the success of previous yolo versions." https://github.com/ultralytics/ultralytics?tab=readme-ov-file, 2023.

Xiaobiao Du is currently pursuing a Ph.D degree at the University of Technology Sydney, Australia. His research interests include 3D vision and generation. He is particularly interested in improving few-shot 3D reconstruction with generative prior.

Haiyang Sun received the BS degree in information and communication engineering from Tsinghua University, Beijing, China, in 2016. He is currently an algorithm expert of LiAuto's Autonomous Driving Team. His research areas include perception algorithms, AIGC, and the neural field.

Ming Lu received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2019. He is currently a Staff Researcher with Intel Labs China, Beijing. His research interests include 3D vision and computer graphics. He is particularly interested in improving the workloads at crucial visual synthesis systems, such as AI + Chips (ISP/Codec/GPU), AIGC, neural field, and large AI models.

Tianqing Zhu received her B.Eng. degree and her M.Eng. degree from Wuhan University, China, in 2000 and 2004, respectively. She also holds a PhD in computer science from Deakin University, Australia (2014). She is currently a professor at city university of Macau. Prior to that, she was a lecturer with the School of Information Technology, Deakin University, associate professor at Sydney University of Technology. Her research interests include privacy preserving, AI security and privacy, and network security.



Xin Yu received the BS degree in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, the PhD degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2015, and the PhD degree from the College of Engineering and Computer Science, Australian National University, Canberra, Australia, in 2019. He is currently a senior lecturer with the University of Queensland. His research interests include computer vision and image processing.