GINA-3D: Learning to Generate Implicit Neural Assets in the Wild

Bokui Shen[1*] Xinchen Yan[2] Charles R. Qi[2] Mahyar Najibi[2] Boyang Deng[1,2†]
Leonidas Guibas[3] Yin Zhou[2] Dragomir Anguelov[2]
[1]Stanford University, [2]Waymo LLC, [3]Google

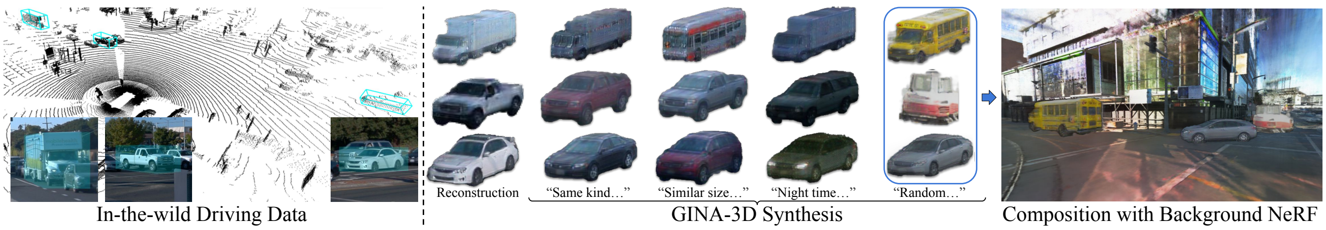


Figure 1. Leveraging in-the-wild data for generative assets modeling embodies a scalable approach for simulation. **GINA-3D** uses real-world driving data to perform various synthesis tasks for realistic 3D implicit neural assets. Left: Multi-sensor observations in the wild. Middle: Asset reconstruction and conditional synthesis. Right: Scene composition with background neural fields [1].

Abstract

Modeling the 3D world from sensor data for simulation is a scalable way of developing testing and validation environments for robotic learning problems such as autonomous driving. However, manually creating or re-creating real-world-like environments is difficult, expensive, and not scalable. Recent generative model techniques have shown promising progress to address such challenges by learning 3D assets using only plentiful 2D images – but still suffer limitations as they leverage either human-curated image datasets or renderings from manually-created synthetic 3D environments. In this paper, we introduce GINA-3D, a generative model that uses real-world driving data from camera and LiDAR sensors to create realistic 3D implicit neural assets of diverse vehicles and pedestrians. Compared to the existing image datasets, the real-world driving setting poses new challenges due to occlusions, lighting-variations and long-tail distributions. GINA-3D tackles these challenges by decoupling representation learning and generative modeling into two stages with a learned tri-plane latent structure, inspired by recent advances in generative modeling of images. To evaluate our approach, we construct a large-scale object-centric dataset containing over 520K images of vehicles and pedestrians from the Waymo Open Dataset, and a new set of 80K images of long-tail instances such as construction equipment, garbage trucks, and cable cars. We compare our model with existing approaches and demonstrate that it achieves state-of-the-art performance in quality and diversity for both generated images and geometries.

*Work done during an internship at Waymo. † Work done at Waymo.

1. Introduction

Learning to perceive, reason, and interact with the 3D world has been a longstanding challenge in the computer vision and robotics community for decades [2–9]. Modern robotic systems [10–16] deployed in the wild are often equipped with multiple sensors (*e.g.* cameras, LiDARs, and Radars) that perceive the 3D environments, followed by an intelligent unit for reasoning and interacting with the complex scene dynamics. End-to-end testing and validating these intelligent agents in the real-world environments are difficult and expensive, especially in safety critical and resource constrained domains like autonomous driving.

On the other hand, the use of simulated data has proliferated over the last few years to train and evaluate the intelligent agents under controlled settings [17–27] in a safe, scalable and verifiable manner. Such developments were fueled by rapid advances in computer graphics, including rendering frameworks [28–30], physical simulation [31, 32] and large-scale open-sourced *asset* repositories [33–39]. A key concern is to create realistic virtual worlds that align in asset content, composition, and behavior with real distributions, so as to give the practitioner confidence that using such simulations for development and verification can transfer to performance in the real world [40–48]. However, manual asset creation faces two major obstacles. First, manual creation of 3D assets requires dedicated efforts from engineers and artists with 3D domain expertise, which is expensive and difficult to scale [26]. Second, real-world distribution contains diverse examples (including interesting rare cases) and is also constantly evolving [49, 50].

Recent developments in the generative 3D modeling offer new perspectives to tackle these aforementioned obstacles, as it allows producing additional realistic but previously unseen examples. A sub-class of these approaches, generative 3D-aware image synthesis [51, 52], holds significant promise since it enables 3D modeling from partial observations (*e.g*. image projections of the 3D object). Moreover, many real-world robotic applications already capture, annotate and update multi-sensor observations at scale. Such data thus offer an accurate, diverse, task-relevant, and up-to-date representation of the real-world distribution, which the generative model can potentially capture. However, existing works use either human-curated image datasets with clean observations [53–58] or renderings from synthetic 3D environments [33, 36]. Scaling generative 3D-aware image synthesis models to the real world faces several challenges, as many factors are entangled in the partial observations. First, bridging the in-the-wild images from a simple prior without 3D structures make the learning difficult. Second, unconstrained occlusions entangle object-of-interest and its surroundings in pixel space, which is hard to disentangle in a purely unsupervised manner. Lastly, the above challenges are compounded by a lack of effort in constructing an asset-centric benchmark for sensor data captured in the wild.

In this work, we introduce a 3D-aware generative transformer for implicit neural asset generation, named GINA-3D (**G**enerative **I**mplicit **N**eural **A**ssets). To tackle the real world challenges, we propose a novel 3D-aware Encoder-Decoder framework with a learned structured prior. Specifically, we embed a tri-plane structure into the latent prior (or *tri-plane latents*) of our generative model, where each entry is parameterized by a discrete representation from a learned codebook [59,60]. The Encoder-Decoder framework is composed of a transformation encoder and a decoder with neural rendering components. To handle unconstrained occlusions, we explicitly disentangle object pixels from its surrounding with an occlusion-aware composition, using pseudo labels from an off-the-shelf segmenation model [61]. Finally, the learned prior of tri-plane latents from a discrete codebook can be used to train conditional latents sampling models [62]. The same codebook can be readily applied to various conditional synthesis tasks, including object scale, class, semantics, and time-of-day.

To evaluate our model, we construct a large-scale object-centric benchmark from multi-sensor driving data captured in the wild. We first extract over 520K images of diverse variations for vehicles and pedestrians from Waymo Open Dataset [14]. We then augment the benchmark with long-tail instances from real-world driving scenes, including rare objects like construction equipment, cable cars, school buses and garbage trucks. We demonstrate through extensive experiments that GINA-3D outperforms the state-of-the-art 3D-aware generative models, measured by image quality, geometry consistency, and geometry diversity. Moreover, we showcase example applications of various conditional synthesis tasks and shape editing results by leveraging the learned 3D-aware codebook. To support future research along this direction, we intend to release the benchmark to support relevant research in the community.

2. Related Work

We discuss the relevant work on generative 3D-aware image synthesis, 3D shape modeling, and applications in autonomous driving.

Generative 3D-aware image synthesis. Learning generative 3D-aware representations from image collections has been increasingly popular for the past decade [63–69]. Early work explored image synthesis from disentangled factors such as learned pose embedding [64,66,69] or compact scene representations [65, 67]. Representing the 3D-structure as a compressed embedding, this line of work approached image synthesis by *upsampling* from the embedding space with a stack of 2D deconvolutional layers. Driven by the progresses in differentiable rendering, there have been efforts [70–73] in baking explicit 3D structures into the generative architectures. These efforts, however, are often confined to a coarse 3D discretization due to memory consumption. Moving beyond explicits, more recent work leverages neural radiance fields to learn implicit 3D-aware structures [51, 52, 74–82] for image synthesis. Schwarz *et al*. [74] introduced the *Generative Radiance Fields (GRAF)* that disentangles the 3D shape, appearance and camera pose of a single object without occlusions. Built on top of *GRAF*, Niemeyer *et al*. [51] proposed the *GIRAFFE* model, which handles scene involving multiple objects by using the compositional 3D scene structure. Notably, the query operation in the volumetric rendering becomes computationally heavy at higher resolutions. To tackle this, Chan *et al*. [52] introduced hybrid explicit-implicit 3D representations with *tri-plane* features *(EG3D)*, which showcases image synthesis at higher resolutions. Concurrently, [83] and [84] pioneer high-resolution unbounded 3D scene generation on ImageNet using *tri-plane* representations, where [84] uses a vector-quantized framework and [83] uses a GAN framework. Our work is designed for applications in autonomous driving sensor simulation with an emphasis on object-centric modeling.

Generative 3D shape modeling. Generative modeling of complete 3D shapes has also been extensively studied, including efforts on synthesizing 3D voxel grids [85–93], point clouds [94–96], surface meshes [97–103], shape primitives [104, 105], and implicit functions or hybrid representations [103, 106–112] using various deep generative models. Shen *et al*. [111] introduced a differentiable explicit surface extraction method called *Deep Marching Tetrahedra (DMTet)* that learns to reconstruct 3D surface meshes with arbitrary topology directly. Built on top of the EG3D [52] *tri-plane* features for image synthesis, Gao *et al*. [103] proposed an extension that is capable of generating textured surface meshes using *DMTet* for geometry generation and *tri-plane* features for texture synthesis. The existing efforts assume access to accurate multi-view silhouettes (often from complete ground-truth 3D shapes) , which does not reflect the real challenges present in data captured in the wild.

Assets modeling in driving simulation. Simulated environment modeling has drawn great attention in the autonomous driving domain. In a nutshell, the problem can be decomposed into asset creation (e.g., dynamic objects and background), scene generation, and rendering. Early work leverages artist-created objects and background assets to build virtual driving environments [18, 20, 113] using classic graphics rendering pipelines. While being able to generate virtual scenes with varying configurations, these methods produce scenes with limited diversity and a significant reality gap. Many recent works explored different aspects of data-driven simulation, including image synthesis [114–117], assets modeling [47, 48, 118–121], scene generation [49, 122, 123], and scene rendering [1, 124–126]. In particular, Chen *et al*. [48] and Zakharov *et al*. [119] performed explicit texture warping or implicit rendering from a single-view observation for each vehicle object. Therefore, their asset reconstruction quality is sensitive to occlusions and bounded by the view angle from a single observation. Building upon these efforts, more recent work including Muller *et al*. [121] and Kundu *et al*. [125] approached object completion with global or instance-specific latent codes, representing each object asset under the Normalized Object Coordinate Space (NOCS). In comparison, the latent codes in our proposed model have 3D tri-plane structures which offers several benefits in learning and applications. More importantly, we can generate previously unseen 3D assets, which is essentially different from object reconstruction.

3. Generative Implicit Neural Assets

We propose **GINA-3D**, a scalable framework to acquire 3D assets from in-the-wild data (Sec. 3.1). Core to our framework is a novel 3D-aware Encoder-Decoder model with a learned structure prior (Sec. 3.2). The learned structure prior can facilitate various downstream applications with an iterative latents sampling model (Sec. 3.3) per application.

3.1. Background.

Given a collection of images containing 3D objects captured in the wild $\mathcal{X} = \{\mathbf{x}\}$ ($\mathbf{x}$ is an image data sample), 3D-aware image synthesis [51, 52, 63–79, 81] aims to learn a distribution of 3D objects. The core idea is to represent each 3D object as a hidden variable $\mathbf{h}$ within a generative model and further leverage a neural rendering module $\texttt{NR}$ to synthesize a sample image at viewpoint $\mathbf{v}$ through $\mathbf{x} = \texttt{NR}(\mathbf{h}, \mathbf{v})$. To model the hidden 3D structure $\mathbf{h}$, the formulation introduces a low-dimensional space where latent variables $\mathbf{z}$ (typically a Gaussian) can sample from and connect $\mathbf{h}$ and $\mathbf{z}$ by a generator $\mathbf{h} = f_\theta(\mathbf{z})$, parameterized by θ.

$$\Pr(\mathbf{x}, \mathbf{z}|\mathbf{v}) = \Pr(\mathbf{x}|\mathbf{z}, \mathbf{v}) \cdot \Pr(\mathbf{z}) \tag{1}$$

The probabilistic formulation is shown in Fig. 2-a, and Eq. 1. Here, $\Pr(\mathbf{x}|\mathbf{z}, \mathbf{v})$ is the conditional probability of the image given the latent variables and viewpoint, where $\Pr(\mathbf{z})$ and $\Pr(\mathbf{v})$ are the prior distributions. As the latent variable $\mathbf{z}$ models the 3D objects, one can sample and extract *assets* for downstream applications. The assets can be either injected into neural representations of scenes [1, 125], or transformed into explicit 3D structures such as textured meshes for traditional renders [20] or geometry-aware compositing [48, 124].

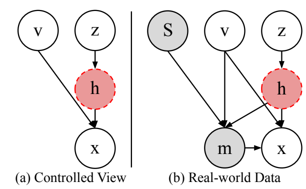


Figure 2. Probabilistic Views.

The challenges in the wild. While human-curated image datasets [53–58] or synthetically generated images with clean background [33, 36, 68, 103] fit into the formulation in Eq. 1, real-world distributions have unconstrained occlusions due to complex object-scene entanglement. For example, a moving vehicle can be easily occluded by another object (*e.g.* traffic cones and cars) in an urban driving environment, which further entangle object and scene in the pixel space. Moreover, environmental lighting and object diversity lead to a more complex underlying distribution.

As illustrated Fig. 2-b and Eq. 2, these challenges yield a new probabilistic formulation that the hidden structure $\mathbf{h}$, surrounding scene $\mathbf{S}$ and viewpoint $\mathbf{v}$ jointly contribute to the occlusion $(\mathbf{m})$ and the visible pixels on the object $\mathbf{x}$ through $\mathbf{x} = \texttt{NR}(\mathbf{h}, \mathbf{v}) \odot \mathbf{m}(\mathbf{S}, \mathbf{v}, \mathbf{h})$.

$$\Pr(\mathbf{x}, \mathbf{z}|\mathbf{v}, \mathbf{S}) = \Pr(\mathbf{x}|\mathbf{z}, \mathbf{v}, \mathbf{S}) \cdot \Pr(\mathbf{z}) \tag{2}$$

Prior art such as GIRAFFE [51] tackles the challenges with two assumptions: (1) the scene is composed of a limited number of same-class foreground objects and a background backdrop $\mathbf{S}$; and (2) the real data distribution can be bridged using an one-pass generator $f_\theta(\mathbf{x}; \mathbf{z}, \mathbf{S}, \mathbf{v})$ (θ is the learned parametrization) conditioned on independently sampled objects $\mathbf{z}$, scene background $\mathbf{S}$ and the camera viewpoint $\mathbf{v}$ (*e.g.* Multi-variate Gaussian distributions with diagonal variance) through adversarial training. Unfortunately, the first assumption barely holds for in-the-wild images with unconstrained foreground occlusions. As shown in Niemeyer *et al*. [51], the second assumption can already introduce artifacts due to disentanglement failures.

Our proposal. We focus on interpreting the visible pixels of the object of interest, as synthesizing objects and scene jointly with a generative model is very challenging. We leverage an auxiliary encoder $E_\phi(\mathbf{x})$ that approximates the posterior $\Pr(\mathbf{z}|\mathbf{x})$ in training the generative model to *reconstruct* the input. This way, we bypass the need to model complex scene and occlusions explicitly, since paired input and output are now available for supervising the auto-encoding style training. Specifically, given an image $\mathbf{x}$ and the corresponding occlusion mask $\mathbf{m}$, our objective is to *reconstruct* the visible pixels of the object on the image through $\hat{\mathbf{x}} \odot \mathbf{m}$ where we have the reconstruction $\hat{\mathbf{x}} = \texttt{NR}(G_\theta(\mathbf{z}), \mathbf{v})$ and latent $\mathbf{z} = E_\phi(\mathbf{x})$, respectively. In practice, we use an off-the-shelf model to obtain the pseudo-labeled object mask as the supervision through $\mathbf{x} \odot \mathbf{m}$. At the inference time, we can discard the auxiliary encoder E_ϕ as our goal is to generate assets from a learned latent distribution (*tri-plane latents* in our case). To facilitate this, we leverage the vector-quantized formulation [59, 60] to learn a codebook $\mathbb{K} := \{z_n\}_{n=1}^K$ of size K and the mapping from a continuous-valued vector to a discrete codebook entry, where each entry follows a K-way categorical distribution.

3.2. 3D Triplane Latents Learning

We explain in details the Encoder-Decoder training framework to learn *tri-plane latents* $\mathbf{z}$ (Fig. 3-left). The framework consists of a 2D-to-3D encoder E_ϕ, learnable codebook quantization $\mathbb{K}$ and a 3D-to-2D decoder G_θ.

E_ϕ: 2D-to-3D Encoder. We adopt Vision Transformer (ViT) [127] as our image feature extractor that maps 16×16 non-overlapping image patches into image tokens of dimension D_{img}. Since the goal is to infer the latent 3D-structure from a 2D image observation, we associate each image token with tokens in the tri-plane latents using cross-attention modules, which have previously shown strong performance in cross-domain and 2D-to-3D information passing [128–131]. The cross-attention module uses a learnable tri-plane positional encoding as query, and image patch tokens as key and value. The module produces tri-plane embeddings $\mathbf{e}^{\text{3D}} = E_\phi(\mathbf{x}) \in \mathbb{R}^{N_Z \times N_Z \times 3 \times D_{\text{tok}}}$, where $D_{\text{tok}} = 32$ and $N_Z = 16$ indicates the dimension of each 3D token and the spatial resolution, respectively.

$\mathbb{K}$: Codebook Quantization for tri-plane latents. Given the continuous tri-plane embedding $\mathbf{e}^{\text{3D}}$, we project it to our K-way categorical prior $\mathbb{K}$ through vector quantization. We apply quantization $\mathbf{q}(\cdot)$ of each spatial code $\mathbf{e}^{\text{3D}}_{ijk} \in \mathbb{R}^{D_{\text{tok}}}$ on the tri-plane embeddings onto its closest entry z_n in the codebook, which gives tri-plane latents $\mathbf{z} = \mathbf{q}(\mathbf{e}^{\text{3D}})$.

$$\mathbf{z}_{ijh} := \left(\underset{z_n, n \in \mathbb{K}}{\operatorname{argmin}} \|\mathbf{e}^{\text{3D}}_{ijk} - z_n\| \right) \in \mathbb{R}^{D_{\text{tok}}} \tag{3}$$

G_θ: 3D-to-2D Decoder with neural rendering. Our decoder takes the tri-plane latents $\mathbf{z}$ as the input and outputs a high-dimensional feature maps $\mathbf{h} \in \mathbb{R}^{N_H \times N_H \times 3 \times D_H}$ used for rendering, where $N_H = 256$ and $D_H = 32$ indicates spatial resolution of the tri-plane feature maps and the feature dimension, respectively. We adopt a token Transformer followed by a Style-based generator [132] as our 3D decoder. The token transformer first produces high-dimensional intermediate features $\hat{\mathbf{z}} \in \mathbb{R}^{N_Z \times N_Z \times 3 \times D_H}$ with an extra `CLS` token using self-attention modules, which are then feed to the Style-based generator for upsampling. We use 4 blocks of weight-modulated convolutional layers, each guided by a mapping network conditioned on the `CLS` token.

Given the feature maps, we use a shallow MLP that takes a 3D point $\mathbf{p}$ and the hidden feature tri-linearly interpolated at the query location $\mathbf{h}(\mathbf{p})$ as input, following [52, 133, 134]. It outputs a density value σ and a view-independent color value $\mathbf{c}$. We perform volume-rendering with the neural radiance field formulation [135].

Training. Our framework builds upon the vector-quantized formulations [59, 60, 62, 136–140] where we focus on token learning in the first stage. Specifically, we extend the VQ-GAN training losses, where the encoder E_ϕ, decoder G_θ and codebook $\mathbb{K}$ are trained jointly with an image discriminator D. As illustrated in Eq. 4, we encourage our Encoder-Decoder model to reconstruct the real image $\mathbf{x}$ with L_2 reconstruction, LPIPS [141], and adversarial loss.

$$\begin{aligned} \mathcal{L}_{\text{RGB}} &= \|(\hat{\mathbf{x}} - \mathbf{x}) \odot \mathbf{m}\|^2 + f^{\text{LPIPS}}(\hat{\mathbf{x}} \odot \mathbf{m}, \mathbf{x} \odot \mathbf{m}) \\ \mathcal{L}_{\text{GAN}} &= [\log D(\mathbf{x}) + \log(1 - D(\hat{\mathbf{x}}))] \end{aligned} \tag{4}$$

To regularize the codebook learning, we apply the latent embedding supervision with a commitment term in Eq. 5, where $\text{sg}[\cdot]$ denotes the stop-gradient operation.

$$\mathcal{L}_{\text{VQ}} = \|\text{sg}[\mathbf{e}^{\text{3D}}] - \mathbf{z}\|_2^2 + \lambda_{\text{commit}}\|\text{sg}[\mathbf{z}] - \mathbf{e}^{\text{3D}}\|_2^2 \tag{5}$$

We additionally regularize the 3D density field in a weakly supervised manner using the rendered aggregated density (alpha value) $\mathbf{x}_\alpha$, encouraging object pixels to have alpha value 1. To make the loss occlusion aware, we further require a pixel lies on the non-object region to have zero density, inspired by Müller *et al*. [121]. This is achieved by restricting the non-object region to cover `sky` or `road` class on the pseudo-labeled segmentations (denoted as $\mathbf{m}_{\text{sky,road}}$).

$$\mathcal{L}_\alpha = \|(\mathbf{x}_\alpha - \mathbf{1}) \odot \mathbf{m}\|^2 + \|\mathbf{x}_{\text{alpha}} \odot \mathbf{m}_{\text{sky,road}}\|^2 \tag{6}$$

To summarize, we optimize the total objective $\mathcal{L}^*$ in Eq. 7.

$$\mathcal{L}^* = \arg \min_{\phi, \theta, \mathcal{Z}} \max_D \mathbb{E}_{\mathbf{x}} \left[\mathcal{L}_{\text{VQ}} + \mathcal{L}_{\text{RGB}} + \mathcal{L}_\alpha + \mathcal{L}_{\text{GAN}}\right] \tag{7}$$

3.3. Iterative Latents Sampling for Neural Assets

Once the first stage training is finished, we can now represent neural assets using the learned *tri-plane latents* and *reconstruct* a collection of assets from image inputs. To generate previously unseen assets with various conditions, we further learn to sample the *tri-plane latents* in the second stage, following the prior works in Generative Transformers [59, 60, 62, 138]. More precisely, we transform the quantized embedding $\mathbf{z} \in \mathbb{R}^{N_Z \times N_Z \times 3 \times D_{\text{tok}}}$ into a discrete sequence $\mathbf{s} \in \{1, ..., K\}^{N_Z \times N_Z \times 3}$, where each element corresponds to the index we select from the codebook $\mathbb{K}$ through $\mathbf{s}_{ijk} = n : \mathbf{z}_{ijk} = z_n$. Following the recent work

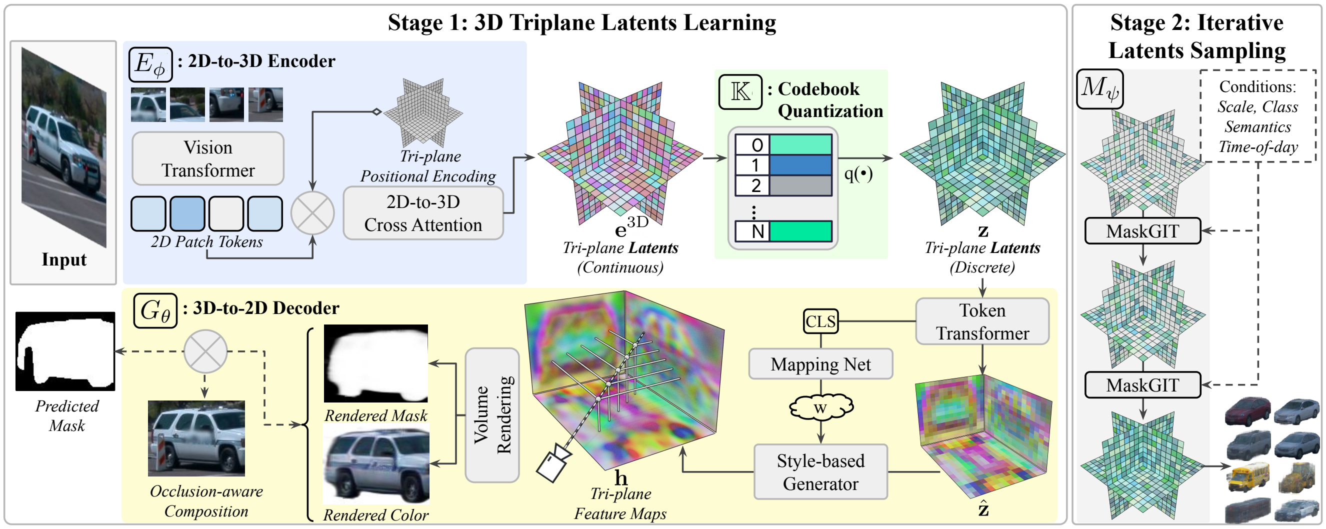


Figure 3. We introduce **GINA-3D**, a 3D-aware generative transformer for implicit neural asset generation. GINA-3D follows a two-stage pipeline, where we learn discrete 3D triplane latents in stage 1 (Sec. 3.2) and iterative latents sampling in stage 2 (Sec. 3.3). In stage 1, an input image is first encoded into continuous tri-plane latents $\mathbf{e}^{3D}$ using a Transformer-based 2D-to-3D Encoder E_ϕ. Then, a learnable codebook $\mathbb{K}$ quantize the latents into discrete latents $\mathbf{z}$. Finally, a 3D-to-2D Decoder G_θ maps $\mathbf{z}$ back to image, using a sequence of Transformer, Style-based Generator and volume rendering. The rendered image is supervised via an occlusion-aware reconstruction loss. In stage 2, we learn iterative latents sampling using MaskGIT [62]. Optional conditional information can be used to perform conditional synthesis. The sampled latents can then be decoded into neural assets using the decoder G_θ learned in stage 1.

MaskGIT [62], we use a bidirectional transformer as our latent generator $M_\psi(\mathbf{z})$ that we learn to iteratively sample the latent sequence (Fig. 3-right). During training, we learn to predict randomly masked latents $\mathbf{s}_{\bar{M}}$ by minimizing the negative log-likelihood of the masked ones.

$$\mathcal{L}_{\text{mask}} = -\mathbb{E}_{\mathbf{s}}[\sum_{\forall ijk:\mathbf{s}_{ijk}=\texttt{[MASK]}} \log \Pr(\mathbf{s}_{ijk}|\mathbf{s}_{\bar{M}})] \tag{8}$$

At inference time, we iteratively generate and refine latents. Starting from all latents as `[MASK]`, we iteratively predict all latents simultaneously but only keep the most confident ones in each step. The remaining ones are assigned as `[MASK]` and the refinement continues. Finally, the sequence $\mathbf{s}$ can be readily mapped back to neural assets by indexing the codebook $\mathbb{K}$ to generate tri-plane latents $\mathbf{z}$ and decoding using G_θ. This iterative approach can be applied to asset variations by selectively masking out tokens of a given instance.

3.4. Expanding Supervision and Conditioning

The two-stage training of GINA-3D is flexible in supervision and conditioning. When we have additional information, we can incorporate it in stage 1 as auxiliary supervision for token learning, or in stage 2 for conditional synthesis.

Unit box vs. Scaled box. Object scale information can serve as an additional input to the tri-linear interpolation on the tri-plane feature maps by rescaling the feature maps to span object bounding box (instead of a unit box).

Semantic feature fields. Various recent works have demonstrated the effectiveness of learning hybrid representations in the neural rendering [142–144] and 2D image synthesis [145]. We can naturally incorporate semantic feature fields in our formulation by computing additional channels in our neural rendering MLP. We precompute DINO-ViT features [146] for each image and learn a semantic feature field to build part correspondence among generated instances.

LiDAR depth supervision. When LiDAR point cloud is available in the data, it can be used as the additional supervision through a reconstruction term between the rendered depth and LiDAR depth.

Conditional synthesis. Last but not the least, additional information support various applications in conditional synthesis. Denoted as $\mathcal{C}$, it can be fed into our latent prior as $M_\psi(\mathbf{s}_{ijk}|\mathbf{s}_{\bar{M}}, \mathcal{C})$. For example, object scale, object class, time-of-day and object semantic embeddings can also serve as c for control over the generation process.

4. Experiments

4.1. Object-centric Benchmark

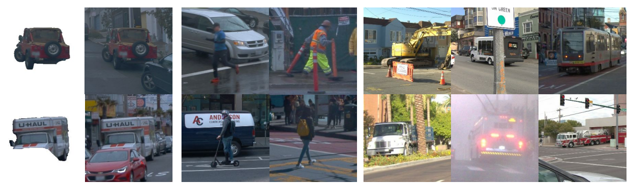
We select the Waymo Open Dataset (WOD) [14] as it is one of the largest and most diverse autonomous driving datasets, containing rich geometric and semantic labels such as object bounding boxes and per-pixel instance masks.

	Images	Unique Instances
WOD-Vehicle	391K	21.0K
WOD-Pedestrian	133K	6.7K
Longtail-Vehicle	80K	3.7K

Table 1. Statistics of our object-centric benchmark.

				Image				Geometry			
				Quality		Semantic Diversity		Quality		Mesh Diversity	
Method	tri-plane $\mathbf{z}$	scaled box	LiDAR	FID↓	Mask FOU↓	COV↑	MMD↓	Cons.↓	Mesh FOU↓	COV↑	MMD↓
GIRAFFE [51]				105.3	43.66	8.24	2.35	15.87	N/A	N/A	N/A
EG3D [52]				137.6	7.40	6.26	2.37	2.38	25.7	3.12	4.70
GINA-3D	×	×	×	147.9	1.85	4.78	2.00	1.55	N/A	1.95	2.43
GINA-3D	✓	×	×	79.0	1.82	19.67	1.52	1.27	11.7	5.75	2.21
GINA-3D	✓	✓	×	60.5	**1.77**	20.68	1.53	1.06	**2.33**	8.69	2.26
GINA-3D	✓	✓	✓	**59.5**	1.80	**25.00**	**1.46**	**0.98**	4.57	**11.42**	**2.17**

Table 2. Quantitative evaluation on the realism and diversity of generated image and geometry (metrics details in Sec. 4.3).



(a) WOD-Vehicle (b) WOD-Pedestrian (c) Longtail-Vehicle

Figure 4. Image samples from our object-centric benchmark.

Specifically, the dataset includes 1,150 driving scenes captured mostly in downtown San Francisco and Phoenix, each consisting of 200 frames of multi-sensor observations. To construct an object-centric benchmark, we propose a *coarse-to-fine* procedure to extract collections of single-view 2D photographs by leveraging 3D object boxes, camera-LiDAR synchronization, and fine-grained 2D panoptic labels. First, we leverage the 3D box annotations to exclude objects beyond certain distances to the surveying vehicle in each data frame (e.g., $40m$ for pedestrians and $60m$ for vehicles, respectively). At a given frame, we project 3D point clouds within each 3D bounding box to the most visible camera and extract the centering patch to build our single-view 2D image collections. Furthermore, we train a Vip-Deeplab model [61, 147] using the 2D panoptic segmentations on the labeled subset and create per-pixel *pseudo-labels* for each camera image on the entire dataset. This allows us to differentiate pixels belonging to the object of interest, background, and occluder (e.g., standing pole in front of a person). We further exclude certain patches where objects are heavily occluded using the 2D panoptic predictions. Even with the filtering criterion applied, we believe that the resulting benchmark is still very challenging due to occlusions, intra-class variations (e.g., truck and sedan), partial observations (e.g., we do not have full 360 degree observations of a single vehicle), and imperfect segmentation. We use the sensor calibrations to compute ray directions for each 2D pixel, taking into account the camera rolling shutter. We repeat the same process to extract vehicles and pedestrians from WOD, and additional longtail vehicles from our Longtail dataset. The proposed object-centric benchmark is one of the largest datasets for generative modeling to date, including diverse and longtail examples in the wild. We intend to release the benchmark to push the frontier of research in this area.

4.2. Implementation Details

GINA-3D. Our encoder takes in images at resolution of 256^2 and renders at 128^2 during training. Our *tri-plane latents* have a resolution of 16^2 with a codebook containing 2048 entries and lookup dimension of 32. We trained our models on 8 Tesla V100 GPUs using Adam optimizer [148], with batch size 32 and 64 in each stage, respectively. We trained stage 1 for 150K steps and stage 2 for 80K steps.

Baselines. We compare against two state-of-the-art methods in the domain, GIRAFFE [51] and EG3D [52], which we train on our dataset at the resolution of 128^2. We noticed that GIRAFFE model trained on full pixels fails to disentangle viewpoints, occlusions and identities. This makes the extraction of the foreground pixels difficult, as the render mask is only defined at the low dimensional resolution 16^2. We instead report the numbers using a model trained by whitening out non-object regions. For EG3D, we observed that training EG3D with unmasked image leads to training collapse, due to the absence of foreground and background modeling. Thus, we trained EG3D under the same setting.

4.3. Evaluations on WOD-Vehicle

We conduct quantitative evaluations in Table. 2 and visualize qualitative results of different model in Fig. 5.

Image Evaluation. For image quality, we calculate Fréchet Inception Distance (FID) [149] between 50K generated images and all available validation images. To better reflect the metric on object completeness, we filter images where its object segmentation mask take up at least 50% of the projected 3D bounding box (Fig.5-right). We additionally measure the completeness of the generated images by Mask Floater-Over-Union (Mask FOU), which is defined as the percentage of unconnected pixels over the rendered object region. To measure the semantic diversity, we compute the Coverage (COV) score and Minimum Matching Distance (MMD) [94] using the CLIP [150] embeddings. COV measures the fraction of CLIP embeddings in the validation set that has matches in the generated set, and MMD measures

Figure 5. Qualitative comparison between GIRAFFE, EG3D and ours with images rendered from a horizontal 30° viewpoint. Both baselines fail to disentangle real-world sensor data. GIRAFFE fails to disentangle rotation in object representation, while both baselines fail to disentangle occlusion and produce incomplete shape. We show samples from occlusion-filtered WOD-Vehicle validation set on the right.

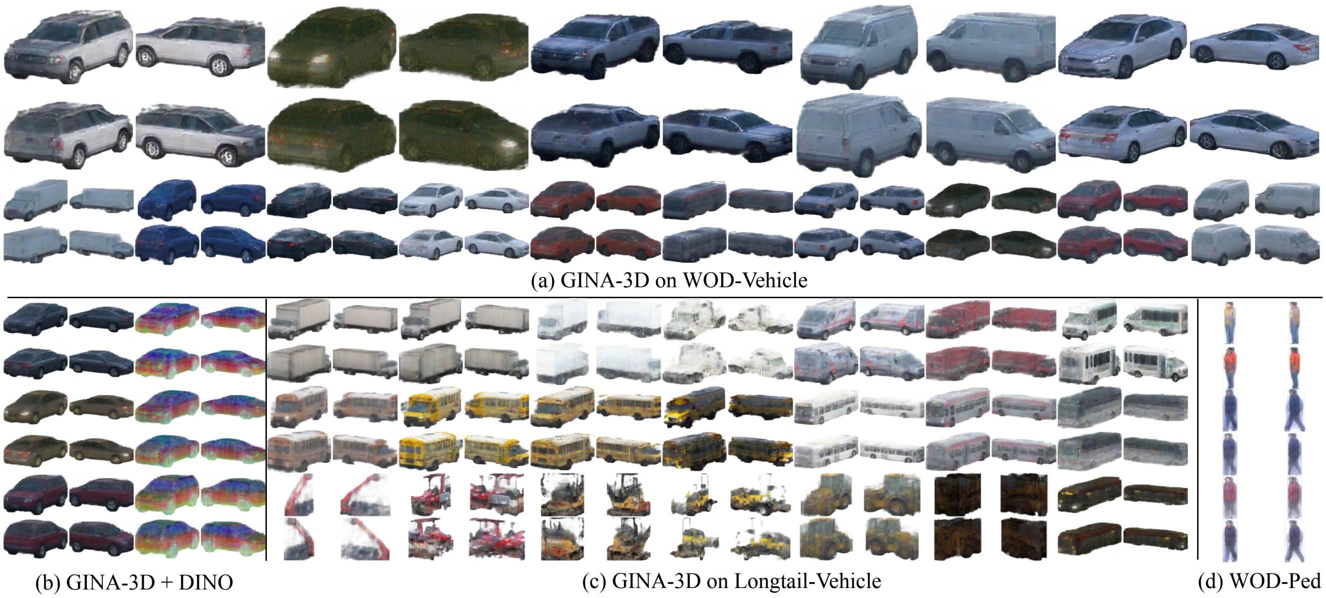


Figure 6. Generation from GINA-3D variants. (a) GINA-3D trained on WOD-Vehicle. (b) GINA-3D with additional DINO feature field generation. (c) GINA-3D trained on Longtail-Vehicle. (d) GINA-3D trained on WOD-Pedestrain.

the distance between each generated embedding to the closest one in the validation. Our model demonstrates significant improvements in FID, image completeness and semantic diversity. Without explicit disentanglement, baselines can hardly handle the real distributions, resulting in artifacts of incomplete shapes (Fig. 5).

Geometry Evaluation. To measure the underlying volume rendering consistency, we follow Or *et al*. [79] and compute the alignment errors between the volume-rendered depth from two viewpoints. We extract the mesh using marching cubes [151] with a density threshold of 10 following EG3D [52]. We measure the completeness by Mesh Floater-Over-Union (Mesh FOU), which is defined as the percentage of the surface area on unconnected mesh pieces over the entire mesh. Since we do not have ground-truth meshes in the real world data, we approximate mesh diversity by measuring between generated meshes and aggregated LiDAR point clouds within a bounding box from the validation set. We measure mesh diversity using the aforementioned COV and MMD with a new distance metric. To account for the incompleteness of LiDAR point clouds, we use a one-way Chamfer distance, which is defined as the mean distance between validation point clouds and their nearest neighbor from a given generated mesh. Our model demonstrates significant improvements in volume rendering consistency, shape completeness and shape diversity.

Augmentation and Ablation. GINA-3D can naturally incorporate additional supervisions when available. We present variations of GINA-3D trained with object scale, LiDAR and DINO [146] supervision. With object scale information available, we normalize tri-plane feature maps with the scale on each dimension. The model trained with rescaled tri-plane resolution yields significant performance boost in both quality and diversity over unit bounding cube, as latents are better utilized. Moreover, we observe that by adding auxiliary L_2 depth supervision from LiDAR, most metrics are improved except Mask and Mesh FOU. While LiDAR provides strong signal to underlying geometry, it also introduces inconsistency on transparent surfaces. We hypothesize that such challenge leads to slightly more floaters, which we leave as future directions to explore. Alternatively, we can learn additional neural semantic fields through 2D-to-3D feature lifting [142]. By only changing the final layer of the `NR` MLP, we can learn an additional view-consistent and

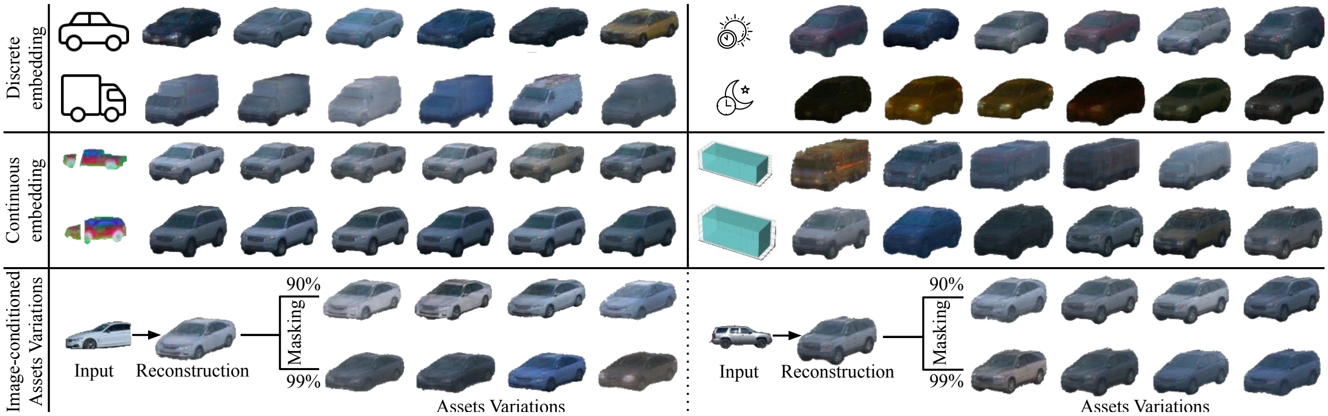


Figure 7. GINA-3D unifies a wide range of asset synthesis tasks, all obtained with the same stage 1 decoder and variations of stage 2 training. Top row: Conditional synthesis using discrete conditions (object classes and time-of-day). 2nd row: Conditional synthesis using continuous conditions (semantic token and object scale). 3rd row: Image-conditioned assets variations by randomizing tri-plane latents.

instance-invariant semantic feature field (Fig. 6-b), which can enable future applications of language-conditioned and part-based editing [8] Finally, we perform ablation studies on the key design of tri-plane latents. If we remove the tri-plane structure and use a MLP-only `NR`, the model fails to capture the diversity of real-world data and results in mode collapse, which generates always a mean car shape.

4.4. Applications

Generating long-tail instance. Our data-driven framework is scalable to new data. We provide results on GINA-3D trained on Longtail-Vehicle and WOD-Ped dataset in Fig. 6-c,d respectively. Without finetuning the architecture on the newly collected data, GINA-3D can readily learn to generate long-tail objects from noisy segmentation masks. As shown in Fig. 6-c, generation results range from trams, truck to construction equipment of various shapes. GINA-3D can also be applied to other categories (e.g. pedestrian, Fig.6-d). Results show moderate shape and texture diversity.

Conditional synthesis. As described in Sec. 3.4, the flexibility of the two-stage approach makes it a promising candidate for conditional asset synthesis. Specifically, we freeze the stage 1 model, and train variations of MaskGIT by passing in different conditions. We provide results for three kinds of conditional synthesis tasks in Fig. 7, namely discrete embeddings (object class, time-of-day), continuous embeddings, and image-conditioned generation. For image-conditioned asset reconstruction and variations, we first infer the latents using the encoder model and then sample asset variations by controlling masking ratio of the reconstructed *tri-plane latents*. The more tokens are masked, the wider the variation range becomes. We provide more details for conditional synthesis in the supplementary material.

4.5. Limitations

Misaligned 3D bounding boxes. As in our WOD-Ped results, misaligned boxes lead to mismatch in pixel space, resulting in blurrier results. Latest methods in ray-based [130] or patch-based [81] learning are promising directions.

Few-shot and transfer learning. Though our data-driven approach achieves reasonable performance by training on Longtail-Vehicle alone, the comparative scarcity of data leads to lower diversity. How to enable few-shot learning or transfer learning remains an open question.

Transcient effects. Direction-dependent effect can be incorporated in our pipeline. We believe modeling material [152] together with LiDAR is an interesting direction.

5. Conclusion

In this work, we presented GINA-3D, a scalable learning framework to synthesize 3D assets from robotic sensors deployed in the wild. Core to our framework is a deep encoder-decoder backbone that learns discrete tri-plane latent variables from partially-observed 2D input pixels. Our backbone is composed of an encoder with cross-attentions, a decoder with tri-plane feature maps, and a neural volumetric rendering module. We further introduce a latent transformer to generate tri-plane latents with various conditions including bounding box size, time of the day, and semantic features. To evaluate our framework, we have established a large-scale object-centric benchmark containing diverse vehicles and pedestrians. Experimental results have demonstrated strong performance on image quality, geometry consistency and geometry diversity over existing methods. To faciliate future research on generative neural assets from in-the-wild data, we intend to release our benchmark to the public.

Acknowledgements: We based our MaskGIT implementation on Chang *et al.* [62]. We thank Huiwen Chang for helpful MaskGIT pointers. We acknowledge the helpful discussions and support from Qichi Yang and James Guo. We thank Mathilde Caron for her DINO implementation and helpful pointers. We based our GIRAFFE baseline on the reimplementation by Kyle Sargent. We thank Golnaz Ghiasi for helpful pointers on segmentation models.

References

[1] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022. 1, 3

[2] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. vis. syst*, 2(3-26):2, 1978. 1

[3] D Man and A Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company, San Francisco*, 1982. 1

[4] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999. 1

[5] Marshall Tappen, William Freeman, and Edward Adelson. Recovering intrinsic images from a single image. *NIPS*, 15, 2002. 1

[6] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pages 577–584. 2005. 1

[7] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 1

[8] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*. IEEE, 2009. 1, 8

[9] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*. Springer, 2010. 1

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[11] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017. 1

[12] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 1

[13] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1, 2, 5

[15] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 1

[16] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[17] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 1

[18] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 1, 3

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[20] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 1, 3

[21] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 1

[22] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*, pages 621–635. Springer, 2018. 1

[23] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 1

[24] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, 126(9):961–972, 2018. 1

[25] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 1

[26] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1

[27] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Claudia Pérez-D'Arpino, Shyamal Buch, Sanjana Srivastava, Lyne Tchapmi, et al. igibson 1.0: a simulation environment for interactive tasks in large realistic scenes. In *IROS*, 2021. 1

[28] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. 1

[29] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, et al. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627*, 2018. 1

[30] Epic Games. Unreal engine. 1

[31] Miles Macklin, Matthias Müller, Nuttapong Chentanez, and Tae-Yong Kim. Unified particle physics for real-time applications. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 1

[32] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016. 1

[33] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 3

[34] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[35] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[36] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M Seitz. Photoshape: Photorealistic materials for large-scale shape collections. *arXiv preprint arXiv:1809.09761*, 2018. 1, 2, 3

[37] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, 2018. 1

[38] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019. 1

[39] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 1

[40] Fereshteh Sadeghi and Sergey Levine. Cad2rl: Real single-image flight without a single real image. *arXiv preprint arXiv:1611.04201*, 2016. 1

[41] Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. Driving policy transfer via modularity and abstraction. *arXiv preprint arXiv:1804.09364*, 2018. 1

[42] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *ICRA*, 2019. 1

[43] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019. 1

[44] Błażej Osiński, Adam Jakubowski, Paweł Zięcina, Piotr Miłoś, Christopher Galias, Silviu Homoceanu, and Henryk Michalewski. Simulation-based reinforcement learning for real-world autonomous driving. In *ICRA*, 2020. 1

[45] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, 2020. 1

[46] Saminda Abeyruwan, Laura Graesser, David B D'Ambrosio, Avi Singh, Anish Shankar, Alex Bewley, and Pannag R Sanketi. i-sim2real: Reinforcement learning of robotic policies in tight human-robot interaction loops. *arXiv preprint arXiv:2207.06572*, 2022. 1

[47] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *CVPR*, 2020. 1, 3

[48] Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchen Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *CVPR*, 2021. 1, 3

[49] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *ICCV*, 2019. 1, 3

[50] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*. IEEE, 2019. 1

[51] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2, 3, 6

[52] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 2, 3, 4, 6, 7

[53] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 3

[54] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015. 2, 3

[55] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 3

[56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 2, 3

[57] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3

[58] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, 2020. 2, 3

[59] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 4

[60] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 4

[61] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, 2021. 2, 6

[62] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 2, 4, 5, 8

[63] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 2, 3

[64] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*. PMLR, 2014. 2, 3

[65] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015. 2, 3

[66] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015. 2, 3

[67] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015. 2, 3

[68] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *NIPS*, 2016. 2, 3

[69] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017. 2, 3

[70] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. In *NeurIPS*, 2018. 2, 3

[71] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, 2019. 2, 3

[72] Thu H Nguyen-Phuoc, Christian Richardt, Long Mai, Yongliang Yang, and Niloy Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. 2020. 2, 3

[73] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR*, 2020. 2, 3

[74] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 2, 3

[75] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, 2021. 2, 3

[76] Zekun Hao, Arun Mallya, Serge Belongie, and Ming-Yu Liu. Gancraft: Unsupervised 3d neural rendering of minecraft worlds. In *ICCV*, 2021. 2, 3

[77] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. 2021. 2, 3

[78] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, 2022. 2, 3

[79] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, 2022. 2, 3, 7

[80] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[81] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. 2022. 2, 3, 8

[82] Kangle Deng, Gengshan Yang, Deva Ramanan, and Jun-Yan Zhu. 3d-aware conditional image synthesis. In *CVPR*, 2023. 2

[83] Ivan Skorokhodov, Aliaksandr Siarohin, Yinghao Xu, Jian Ren, Hsin-Ying Lee, Peter Wonka, and Sergey Tulyakov. 3d generation on imagenet. In *International Conference on Learning Representations*. 2

[84] Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Srinivasan, Jiajun Wu, and Deqing Sun. Vq3d: Learning a 3d-aware generative model on imagenet. *arXiv preprint arXiv:2302.06833*, 2023. 2

[85] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2

[86] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. 2016. 2

[87] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*. Springer, 2016. 2

[88] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *3DV*, 2017. 2

[89] Edward J Smith and David Meger. Improved adversarial systems for 3d object generation and reconstruction. In *CoRL*. PMLR, 2017. 2

[90] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *ICCV*, 2019. 2

[91] Sebastian Lunz, Yingzhen Li, Andrew Fitzgibbon, and Nate Kushman. Inverse graphics gan: Learning to generate 3d shapes from unstructured 2d data. *arXiv preprint arXiv:2002.12674*, 2020. 2

[92] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 2

[93] Moritz Ibing, Gregor Kobsik, and Leif Kobbelt. Octree transformer: Autoregressive 3d shape generation on hierarchically structured sequences. *arXiv preprint arXiv:2111.12480*, 2021. 2

[94] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*. PMLR, 2018. 2, 6

[95] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *ICCV*, 2019. 2

[96] Kaichun Mo, He Wang, Xinchen Yan, and Leonidas Guibas. PT2PC: Learning to generate 3d point cloud shapes from part tree conditions. 2020. 2

[97] Ayan Sinha, Asim Unmesh, Qixing Huang, and Karthik Ramani. Surfnet: Generating 3d shape surfaces using deep residual networks. In *CVPR*, 2017. 2

[98] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018. 2

[99] Dario Pavllo, Graham Spinks, Thomas Hofmann, Marie-Francine Moens, and Aurelien Lucchi. Convolutional generation of textured 3d meshes. *NeurIPS*, 2020. 2

[100] Dario Pavllo, Jonas Kohler, Thomas Hofmann, and Aurelien Lucchi. Learning generative models of textured 3d meshes from real-world images. In *ICCV*, 2021. 2

[101] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. *NeurIPS*, 2019. 2

[102] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *ICML*. PMLR, 2020. 2

[103] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In *NeurIPS*, 2022. 2, 3

[104] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *ACM Transactions on Graphics (TOG), Siggraph Asia 2019*, 38(6):Article 242, 2019. 2

[105] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 2

[106] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *NeurIPS*, 2019. 2

[107] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[108] Andrew Luo, Tianqin Li, Wen-Hao Zhang, and Tai Sing Lee. Surfgen: Adversarial 3d shape synthesis with explicit surface discriminators. In *ICCV*, 2021. 2

[109] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[110] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *CVPR*, 2022. 2

[111] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *NeurIPS*, 2021. 2

[112] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *CVPR*, 2022. 2

[113] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*. Springer, 2016. 3

[114] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. *NeurIPS*, 2018. 3

[115] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 3

[116] Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019. 3

[117] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. In *NeurIPS*, 2020. 3

[118] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *NeurIPS*, 2021. 3

[119] Sergey Zakharov, Rares Andrei Ambrus, Vitor Campagnolo Guizilini, Dennis Park, Wadim Kehl, Fredo Durand, Joshua B Tenenbaum, Vincent Sitzmann, Jiajun Wu, and Adrien Gaidon. Single-shot scene reconstruction. In *CoRL*, 2021. 3

[120] Tom Monnier, Matthew Fisher, Alexei A Efros, and Mathieu Aubry. Share with thy neighbors: Single-view reconstruction by cross-instance consistency. In *ECCV*, 2022. 3

[121] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *CVPR*, 2022. 3, 4

[122] Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *ECCV*. Springer, 2020. 3

[123] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *CVPR*, 2021. 3

[124] Zhenpei Yang, Yuning Chai, Dragomir Anguelov, Yin Zhou, Pei Sun, Dumitru Erhan, Sean Rafferty, and Henrik Kretzschmar. Surfelgan: Synthesizing realistic sensor data for autonomous driving. In *CVPR*, 2020. 3

[125] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *CVPR*, 2022. 3

[126] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 3

[127] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[128] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *ICCV*, 2021. 4

[129] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 4

[130] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, 2022. 4, 8

[131] Daniel Rebain, Mark J Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields. *arXiv preprint arXiv:2209.10684*, 2022. 4

[132] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 4

[133] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*. Springer, 2020. 4

[134] Bokui Shen, Zhenyu Jiang, Christopher Choy, Leonidas J Guibas, Silvio Savarese, Anima Anandkumar, and Yuke Zhu. Acid: Action-conditional implicit visual dynamics for deformable object manipulation. 2022. 4

[135] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4

[136] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 4

[137] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*. PMLR, 2020. 4

[138] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In *ICLR*, 2021. 4

[139] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 4

[140] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 4

[141] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4

[142] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *arXiv preprint arXiv:2205.15585*, 2022. 5, 7

[143] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *arXiv preprint arXiv:2209.03494*, 2022. 5

[144] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *arXiv preprint arXiv:2210.15947*, 2022. 5

[145] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, 2021. 5

[146] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5, 7

[147] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Yukun Zhu, Liang-Chieh Chen, Henrik Kretzschmar, and Dragomir Anguelov. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022. 6

[148] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[149] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017. 6

[150] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 6

[151] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 7

[152] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 8