VQ3D: Learning a 3D-Aware Generative Model on ImageNet

Kyle Sargent[1] Jing Yu Koh[2*] Han Zhang[3] Huiwen Chang[4*] Charles Herrmann[3]
Pratul Srinivasan[3] Jiajun Wu[1] Deqing Sun[3]
[1]Stanford University [2]Carnegie Mellon University [3]Google Research [4]OpenAI

Abstract

Recent work has shown the possibility of training generative models of 3D content from 2D image collections on small datasets corresponding to a single object class, such as human faces, animal faces, or cars. However, these models struggle on larger, more complex datasets. To model diverse and unconstrained image collections such as ImageNet, we present VQ3D, which introduces a NeRF-based decoder into a two-stage vector-quantized autoencoder. Our Stage 1 allows for the reconstruction of an input image and the ability to change the camera position around the image, and our Stage 2 allows for the generation of new 3D scenes. VQ3D is capable of generating and reconstructing 3D-aware images from the 1000-class ImageNet dataset of 1.2 million training images, and achieves a competitive ImageNet generation FID score of 16.8. Our project webpage is at this url.

1. Introduction

3D assets are an important part of popular media formats such as video games, movies, and computer graphics. Since 3D content can be time-consuming to create by hand, automatically generating 3D content using machine learning is an active area of research. While machine learning techniques benefit from training on large amounts of data, existing 3D datasets have noisy labels and are orders of magnitude smaller than those of 2D images.

To circumvent the limitations of 3D datasets, recent GAN-based methods have explored learning generative models of 3D scenes from images with limited or no 3D labels [26, 6, 18, 25]. These GAN-based approaches demonstrate the promise of learning 3D representations from 2D data. However, GANs are unstable and challenging to scale to large diverse datasets [3, 35]. Because of these issues, recent 3D-aware GANs mostly focus on single-class datasets, such as human faces [20], animal faces [8], or cars [48].

In order to move beyond single-class generation, we draw inspiration from recent advances in 2D image generation, where other formulations such as text-to-image generation models [51, 33, 28] and two-stage image models [14, 50] have begun to achieve impressive results on very large and diverse image collections. The most recent state-of-the-art 2D generative models leverage diffusion or vector quantization rather than GANs to scale well to large datasets. In particular, the two-stage vector quantization approach, being a likelihood method, can model more diverse modes and is more stable during training. This motivates us to explore vector quantization as an alternative to the popular GAN-based methods for 3D-aware generative models.

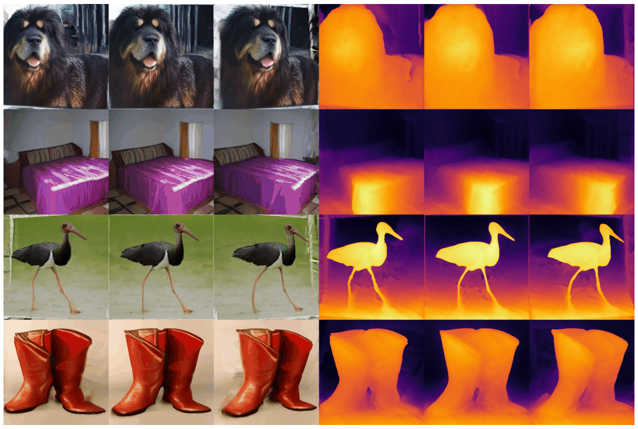

Figure 1: 3D-aware images generated by VQ3D on ImageNet. Please see supplemental materials for video results.

In this paper, we propose VQ3D, a strong 3D-aware generative model that can be learnt from large and diverse 2D image collections, such as ImageNet [9]. To encourage stability and higher reconstruction quality, we forgo GAN-based [16] approaches [25, 5, 6, 26, 18], in favor of the two-stage autoencoder formulation of VQGAN [14] and ViT-VQGAN [50]. We build on this formulation with several novel architectural components and losses, and show through ablations that they are necessary for good performance and 3D-awareness. We learn 3D geometry by introducing a conditional NeRF decoder and modified triplane representation which can handle unbounded scenes, and training with a novel loss formulation which encourages high-quality geom-

[*] Work completed at Google Research.

etry and novel views.

Our formulation has two advantages that ensure its ability to scale and correctly model ImageNet. First, we separate the training into two stages (reconstruction and generation). This enables us to directly supervise the first stage training via a novel depth loss that uses pseudo-GT depth. Note, this is possible because our conditional NeRF decoder in the first stage learns to both reconstruct the input and predict the depth of each image.

Second, our two-stage formulation is simpler and more reliable than existing techniques for training 3D-aware generative models. Our formulation does not use progressive growing [5, 6], a neural upsampler [6, 5, 26], pose conditioning [6, 42], or patch-wise discriminators [36, 42], yet still learns meaningful 3D representations. Previous work [2, 35] found that 2D GANs cannot easily scale up to large diverse datasets (*e.g*., ImageNet) and significant innovations in training techniques are needed. Despite an exhaustive hyperparameter search, we were unable to scale existing 3D GAN baselines to ImageNet, and future training innovations are needed to make 3D GANs work on ImageNet. By contrast, our two-stage formulation scales to ImageNet stably like prior two-stage architectures [50, 14], and also achieves comparable or superior performance to existing 3D GAN baselines on simpler datasets such as CompCars [48].

We verify that baseline 3D-aware GAN methods [6, 18, 5, 26], while working well on single-object datasets, fail to learn good generative models for ImageNet. Compared to the best existing 3D-aware baseline, VQ3D attains a 75.9% relative improvement on FID scores for 3D-aware ImageNet images (69.8 for StyleNeRF [18] to 16.8 for VQ3D).

In summary, we make the following three contributions:

- We present a novel 3D-aware generative model that can be trained on large and diverse 2D image collections. Whereas all previous methods are GANs, we are the first to show that a two-stage VQ formulation can work for 3D-aware generative models. Our two-stage inherits the stability of prior VQ formulations and works reliably on both single-class and highly diverse datasets. Our formulation also allows the use of pseudo-depth supervision in the first stage.

- We obtain state-of-the-art generation results on ImageNet, demonstrating that our 3D-aware generative model is capable of fitting a dataset at the scale and diversity of ImageNet. Our model significantly outperforms the next best baseline.

- The Stage 1 of our model enables 3D-aware image editing and manipulation. One forward pass through our network converts a single RGB image into a manipulable NeRF, without relying on an expensive inversion optimization used in prior work [5, 6].

2. Related Work

3D-aware generative models. Several recent papers tackle the task of modeling 3D-aware generation, primarily through the GAN framework [16]. HoloGAN [25] learns perspective projection and rendering of 3D features, and applies 3D rigid-body transforms to generate new images from different poses. More recently, several papers use NeRF [24] as the 3D backbone [26, 5, 18, 47], which allows the 3D scene to be defined as a 3D volume parameterized by an MLP. EG3D [6] proposes a hybrid triplane representation which scales well with resolution, and enables greater generation detail. Disentangled3D [43] learns a 3D-aware GAN from monocular images with disentangled geometry, appearance, and pose. Pix2NeRF [4] proposes a method for unsupervised learning of neural representations with a shared pose prior, which enables rendering of novel views from a single input image. GRAF [36] and EpiGRAF [42] train 3D GANs via patch-wise representations to save on the expense of volume rendering. GRAM [10] proposes learning a set of implicit surfaces, shared for the training object category. At inference time, images are generated by accumulating the radiance along each ray using ray-surface intersections as samples. All these works are GANs designed to be trained on single-class image collections, whereas we propose a two-stage autoencoder architecture which can be trained on diverse datasets such as ImageNet.

Conditional NeRF and other 3D representations. Recent work has focused on the appropriate way to condition NeRF to achieve maximum expressiveness. GIRAFFE [26] demonstrated success with the "conditioning-by-concatenation" approach [40], in which the scene's latent codes are fed into the first layer of the NeRF MLP and not thereafter. Other work such as pi-GAN [5] transforms the latent code into a vector of frequencies and phase shifts for each layer of a SIREN [39]. Other work has used hypernetworks [40, 38] to parameterize 3D representations, and MetaSDF [38] showed that many forms of conditioning are special cases of hypernetworks. Our model can be seen as a conditional NeRF. We show that our novel decoder architecture, consisting of a ViT [12] and contracted triplane representation, is powerful enough to encode and reconstruct all of ImageNet. Given a single image, we show that in a single forward pass and without any optimization, our model can create a NeRF of an input RGB image with reasonable reconstruction at the main view and plausible novel views.

Quantization models. Image quantization is a powerful paradigm used in recent state-of-the-art generative models. In this setup, an image is encoded into a discrete latent representation [45], which improves generation quality when paired with an autoregressive generative prior (most often a transformer [46]). This has led to impressive results in image generation [15, 50, 23], text-to-image generation [29, 11, 51], and other tasks. Recent image quantization models improve reconstruction quality by introducing adversarial losses [15], using vision transformer encoders and decoders (ViT) [12, 50] as both encoder and decoder, representing discrete codes as a stacked map [23], and more. Such quantization architectures typically use powerful CNNs [14] or ViT [50] encoders and decoders; ViT and CNN-based architectures show good performance reconstructing large datasets. However, these architectures encode and decode 2D feature maps and so are not inherently 3D-aware. in this paper, we show that our 3D-aware NeRF-based decoder can also work well in the quantization framework. It has the capacity to encode and reconstruct ImageNet, and also learns a discrete latent codebook that can be used to train a powerful fully generative Stage 2 model.

Single-view 3D reconstruction and novel view synthesis. Various approaches for 3D reconstruction or novel view synthesis in the context of generative or auto-encoder models have been proposed. Kato et. al [21] propose an adversarial training scheme using two discriminators for single-view 3D reconstruction. Their scheme inspires our use of two discriminators for similar reasons. However, their model cannot sample totally new scenes. More recently, uORF [49] uses NeRFs as 3D object representations to enable 3D scene decomposition. uORF represents a 3D scene as a composition of an object radiance field for each object, and a background radiance field for the remainder of the scene. This enables re-rendering and editing of 3D scenes from an input image. However, uORF also cannot sample new scenes, and moreover requires multi-view training datasets.

In the domain of novel scene generation, Generative Query Networks (GQN) [13] use CNNs to represent and generate scenes. GQNs can imagine and re-render scenes from novel viewpoints, but due to the usage of CNNs, do not explicitly embed 3D geometry or have any guarantees of scene consistency. NeRF-VAE [22] proposes a VAE representation which models multiple scenes. Unlike GQNs, which have no 3D prior, NeRF-VAE uses NeRF to achieve 3D consistency. However, it relies on multi-view training data. LOLNeRF [32] learns a generative model of 3D face images but requires a keypoint estimator and the auto-decoder formulation requires an optimization to be applied to examples outside its training set. By contrast, our method can be applied to single RGB images and requires only 2D training data and an off-the-shelf depth estimator during training.

Concurrent ImageNet-focused works Two concurrent recent works, 3DGP [41] and IVID [19], have achieved strong ImageNet FID, but with distinct approaches to VQ3D. 3DGP leverages GANs, while IVID leverages diffusion models. It is our hope that having multiple avenues to reach good ImageNet performance (VQ, GAN, and diffusion-based) will be beneficial to the community. Interestingly, 3DGP, IVID, and VQ3D all use off-the-shelf monodepth estimators to achieve good ImageNet performance. It is an open question whether 3D generative models can be learned on ImageNet without depth estimators. We urge the readers to read these papers to have a more comprehensive view of this active research area.

3. Model

3.1. Overview of VQ3D

Two-stage VQ frameworks such as VQGAN [14], ViT-VQGAN [50], and Parti [51] have demonstrated compelling performance on datasets at the scale of ImageNet or even larger. All prior work in 3D-aware generative models is GAN-based, but the strong performance of two-stage VQ frameworks on diverse data motivates us to apply them to the learning of 3D-aware generative models.

Our model is a vector-quantized autoencoder [50, 14, 51], which is trained in two stages. Stage 1 of our model consists of an encoder and decoder. The encoder encodes RGB images into a learned latent codebook, and the decoder reconstructs them. A diagram of the inputs, outputs, and architecture of the first stage is given in the top of Figure 2. The encoder of our first stage is a ViT similar to VIM [50], but the decoder is a conditional NeRF, which allows us to introduce 3D-awareness. The first stage is trained end-to-end by encoding and reconstructing RGB training images while minimizing reconstruction and adversarial losses. After training, the first stage can be used to encode unseen single RGB images and then reconstruct them in 3D, which enables novel view synthesis, image editing and manipulations.

Stage 2 is an autoregressive transformer which predicts sequences of latent tokens. A diagram of the inputs, outputs, and architecture is shown in the bottom of Figure 2. It is trained on sequences of latent codes produced by our Stage 1 encoder. After training, the autoregressive transformer can be used to generate totally new 3D images by first sampling a sequence of latent tokens and then applying our NeRF-based decoder. Importantly, our Stage 2 model inherits the properties optimized in Stage 1, so the fully generated images have high-quality geometry and plausible novel views.

3.2. Training

We now provide additional training details for the two stages of our model.

Stage 1. The goal of the first stage is to learn a model which can compress image pixels into a sequence of discrete indices corresponding to a learnt latent codebook [50, 14]. Since we desire our model to be 3D-aware, we impose several additional criteria:

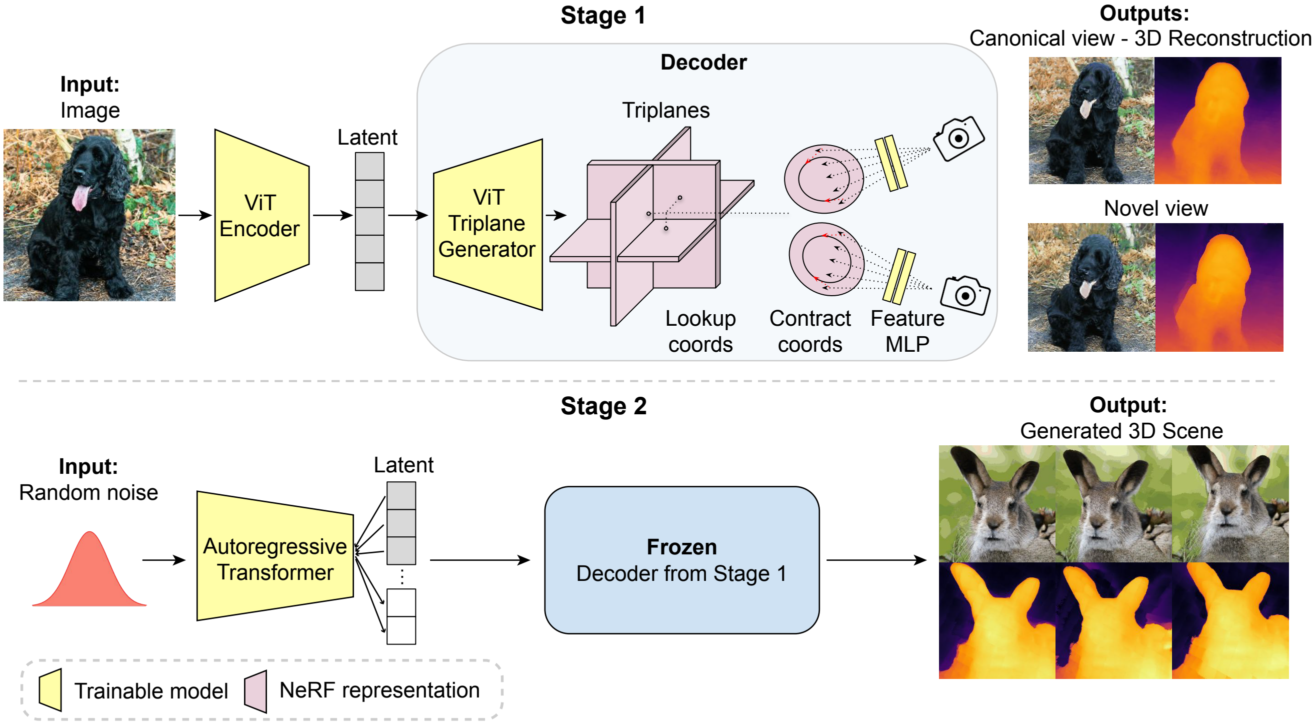

Figure 2: VQ3D model architecture. We propose a novel NeRF-based decoder that makes both stages of VQ3D 3D-aware.

1. **Good reconstruction at the canonical view.** On ImageNet, ground truth camera extrinsics are unknown and probably not even well-defined due to the presence of deformable and ambiguous object categories and scenes without salient objects. Therefore, we fix the same identical 'canonical view' for all images. Our criterion is that our conditional NeRF-based autoencoder should successfully reconstruct every image in the dataset from this view.

2. **Reasonable novel views.** We expect that images decoded at novel views within a specified range of the canonical view (referred to as the "sampling radius") will have similar quality to images decoded at the canonical view.

3. **Correct depth.** The depth of the reconstructed scene viewed from the canonical view should correspond to the GT depth of the image up to scale and shift.

We enforce these criteria by introducing several auxiliary models and losses. Our novel losses are diagrammed in Figure 3. To enforce good reconstruction at the canonical view, we train with a combination of the MSE loss $\mathcal{L}_{\text{MSE}}$, the perceptual loss $\mathcal{L}_{\text{percep}}$, and the Logit-Laplace loss $\mathcal{L}_{\text{log-lap}}$, following prior work [50].

To enforce reasonable novel views, we leverage a main and auxiliary discriminator, similar to Kato and Harada [21]. The first discriminator distinguishes between real and reconstructed images at the canonical viewpoint, while the second distinguishes between reconstructed images at the canonical viewpoint and novel views. In this way, the model cannot allocate all its capacity to reconstructing images at the canonical viewpoint without also having high-quality novel views. As prior work noted [21], the generator may slightly corrupt the main view in order to collaborate with the novel view branch to fool the discriminator; thus, we add a stop-grad between the main view and the novel view discriminator. We sample novel views during training uniformly in a disc tangent to a sphere at the canonical camera pose. We use the non-saturating GAN objective $\mathcal{L}_{\text{gan}}$ [16] for both discriminators. We additionally concatenate the predicted depth as input to the auxiliary discriminator to ensure the distribution of depths does not change depending on the camera viewpoint.

To enforce correct depth and geometry, we supervise the NeRF depth with pseudo-GT geometry at the main viewpoint. We employ the pretrained depth prediction transformer model DPT [30] which produces pseudo-GT inverse depth estimates for the images in our training datasets. Thus, our model is limited to some extent by the quality of the depth estimator chosen. [31] proposed a shift- and scale-invariant l_2 loss for training monocular depth estimation in which the shift and scale are determined by solving a closed-form least squares alignment with the GT depth. We propose a novel formulation of this shift- and scale- invariant loss adapted to the NeRF setting, in which we supervise the weight of every sample along each ray rather than the accumulated depth. For a given image, let $i \in \{1...N\}$ and $k \in \{1...L\}$ be indices which range over the image plane and ray samples respectively, let D_{ik} be the pointwise inverse depths of the NeRF sample locations, let W_{ik} be corresponding NeRF weights from volumetric rendering [24], and let d_i be the pseudo-GT depth from DPT. Then we define s^*, t^* to be the closed-form solution of the weighted least squares problem:

$$s^*, t^* = \arg\min_{s,t} \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{L} W_{ik}(sD_{ik} + t - d_i)^2. \quad (1)$$

We set our depth loss to be the weighted scale- and shift-invariant loss:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{d=1}^{L} W_{ik}(s^* D_{ik} + t^* - d_i)^2. \quad (2)$$

Assuming the weight sum to 1 along each ray, this loss is minimized when the NeRF allocates zero weight to all but one sample location along each ray, and the expectation with respect to the weights of the inverse depth is equal to the GT inverse depth map up to a scale and shift. In this way, it functions similarly to the distortion loss [1] by penalizing weight distributions which are too spread out, but also encourages the weights to be concentrated near the GT surface. Importantly, this formulation still allows for more than one surface along each ray and thus for occlusion and disocclusion, because the penalty is applied to the volumetric rendering weights and not the predicted density. We find this depth loss formulation to be critical for good performance. In particular, supervising the accumulated inverse depth rather than the pointwise inverse depths leads to poor performance, and we provide an ablation of this and other design choices in the supplementary material.

We additionally introduce two penalties on the scale determined by this alignment:

$$\mathcal{L}_{\text{scale}} = \max(0, -s^*) + \lambda_{\text{smallscale}} \max(s^* - 1, 0). \quad (3)$$

The first term is a small penalty to prevent the sign of the inverse depth scale from flipping negative. The second term is a penalty preventing the inverse depth maps from becoming too flat, which encourages perceptually pleasing novel views. $\lambda_{\text{smallscale}}$ gives the relative strength of the second term in the scale loss. We use the same vector-quantization loss $\mathcal{L}_{\text{vq}}$ [50], and the distortion and interlevel losses of Mip-NeRF360 [1], given by $\mathcal{L}_{\text{distort}}$, $\mathcal{L}_{\text{interlevel}}$. The loss for our Stage 1 model with associated weights λ is thus:

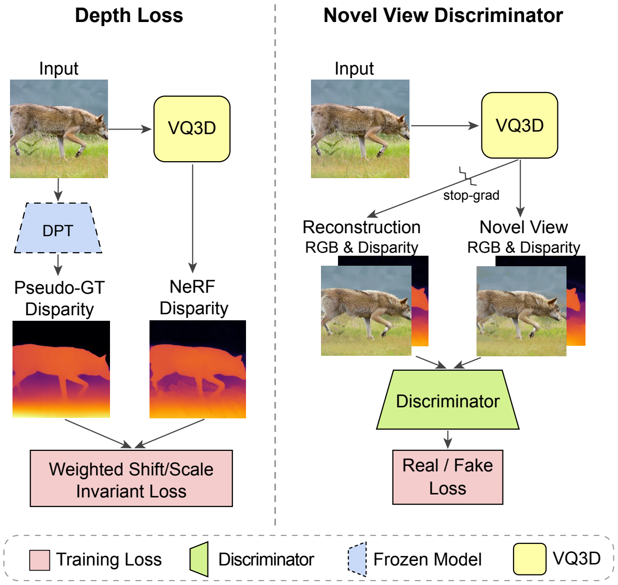

Figure 3: Diagram of the key novel losses in Stage 1 optimization. The depth losses enforces correct geometry of the reconstructed scene, while the novel view discriminator enforces reasonable novel views as well as inpainting and outpainting.

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{percep}} \cdot \mathcal{L}_{\text{percep}} + \lambda_{\text{log-lap}} \cdot \mathcal{L}_{\text{log-lap}} + \\ \lambda_{\text{gan}} \cdot \mathcal{L}_{\text{gan}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{depth}} + \lambda_{\text{scale}} \cdot \mathcal{L}_{\text{scale}} + \\ \lambda_{\text{vq}} \cdot \mathcal{L}_{\text{vq}} + \lambda_{\text{distort}} \cdot \mathcal{L}_{\text{distort}} + \lambda_{\text{interlevel}} \cdot \mathcal{L}_{\text{interlevel}}. \quad (4)\end{aligned}$$

The exact settings of the loss weights are given in the supplementary material.

Stage 2. The goal of Stage 2 is to learn an autoregressive model over the discrete encodings produced by the Stage 1 encoder, so that completely new 3D scenes can be generated. Our Stage 2 transformer and training details follow ViT-VQGAN [50]. We verify experimentally that our fully generative Stage 2 model inherits the properties optimized in Stage 1; namely, 3D-consistent novel views and high quality geometry. We also apply top-k and top-p filtering [15].

3.3. Architecture

A full architecture diagram is shown in Figure 2. Similar to ViT-VQGAN [50], we leverage the powerful vision transformer [12] architecture in both the encoder and decoder. Different from ViT-VQGAN [50], which is trained on 2D images, we utilize a novel decoder with 3D inductive bias to facilitate the learning of 3D representations. We now give an overview of the individual components of our architecture.

Encoder and triplane decoder. For the encoder, we use a ViT-S model. For the decoder, we use a ViT-L model to decode the latent codes into three triplanes of size 512x512 with feature dimension 32.

Contracted triplane representation & NeRF MLP. We must reconstruct and generate potentially unbounded ImageNet scenes, but we are motivated to leverage the powerful triplane representation [6], Therefore, we propose an adapted triplane representation borrowing from both [6] and [1]. We apply the contraction function of MipNeRF360 to bound coordinates within the triplanes before looking up their values, and use the linear-in-inverse depth sampling scheme with separate proposal and NeRF MLP. The MLPs convert interpolated triplane features to density and, in the case of the NeRF MLP, RGB color. Similar to [6], our MLPs are lightweight, with 2 layers and 32 hidden units each; unlike [6], we directly render RGB color rather than using a neural upsampler, as we found neural upsampling to be a source of myriad and confusing artifacts not fixable via dual discriminators [6] or consistency losses [18].

Autoregressive transformer. We train a transformer [46] to autoregressively predict the next image token. We follow the hyperparameters in the base model of VIM [50]. For ImageNet, we train a class-conditional model, and for other datasets we train unconditional generative models.

4. Experiments

4.1. Main results

We study the performance of our method and the baseline methods on ImageNet, a standard benchmark for 2D image generation which consists of 1.28M images of 1000 object classes. For our main results on ImageNet, we training for the longest possible time and use the most optimal top-p and top-k sampling parameters. For our later analysis experiments, we use a consistent Stage 1 and Stage 2 step across each study, and do not using top-p or top-k sampling unless for ablation.

We compare against pi-GAN [5], GIRAFFE [26], EG3D [6], and StyleNeRF [18]. We re-implemented pi-GAN and GIRAFFE using our internal framework, and ran the provided code for EG3D and StyleNeRF. Since ImageNet does not have GT poses and pseudo-GT poses are not possible to compute, we disable generator and discriminator pose conditioning for EG3D and sample from a pre-defined pose distribution. We extensively tune the strongest baselines, StyleNeRF and EG3D, on ImageNet, and report the best results from all runs. All results from our hyperparameter sweeps are given in the supplementary material.

Our main results for generation on ImageNet are given in Table 1. Notably, our FID score on ImageNet is the best by a wide margin. We show generated examples from our method and the benchmarks in Figure 4 and note our method generates superior samples.

Generation	Type	FID ↓	IS ↑	Depth Acc. ↓
StyleGAN-XL [35]	2D	2.30	265.1	-
ViT-VQGAN [50]	2D	4.17	175.1	-
pi-GAN [5]	3D	101.4	9.7	1.41
GIRAFFE [26]	3D	132.1	9.2	1.78
StyleNeRF [18]	3D	69.8	15.5	1.96
EG3D [6]	3D	82.2	13.3	1.93
VQ3D (Ours)	3D	**16.8**	**82.9**	**0.13**

Table 1: FID scores of generative models on ImageNet. We set a new state of the art on ImageNet with a more than fourfold improvement over the next best 3D-aware baseline. 2D methods included for comparison purposes.

In addition to generating high quality scenes, Stage 1 of our method can also be used for single-view 3D reconstruction and manipulation. Figure 5 shows single RGB images reconstructed by our Stage 1 with estimated geometry. Our network performs well at reconstruction and needs only a single forward pass to compute a NeRF for an input image, unlike prior work [6, 5] which requires an inversion optimization. Moreover, the reconstructed NeRFs can be manipulated, for instance to render novel views. We show examples of novel views in Figure 6.

4.2. Analysis and ablations

We perform analysis on the use of depth losses and learning of geometry, both for our model and the baseline methods. We then conduct an in-depth ablation study on the design choices of VQ3D. We additionally compare VQ3D against a combination of a 2D GAN and novel view synthesis model. Finally, we show results for various settings of top-p and top-k sampling.

One potential concern may be that the use of pseudo-GT depth limits the comparability of our technique with the baseline GAN methods. We address this concern by analyzing both the FID score and the depth accuracy metric used in [37, 6]. This metric is defined as the mean- and variance-normalized MSE between the NeRF depth and the predicted depth of the generated image. Table 2 gives the result for generative models with and without depth losses. For the GAN methods, we find that our pointwise inverse depth loss works poorly but directly supervising with depth accuracy seems to improve geometry, except for StyleNeRF for which no depth losses appear to work. For our method, we show the Stage 2 performance with and without our novel pointwise weighted depth loss. While performance on the depth accuracy metric can improve when various depth losses are incorporated training, the effect on FID is negligible, suggesting that incorporating pseudo-GT depth is unlikely to meaningfully improve the FID for the baseline

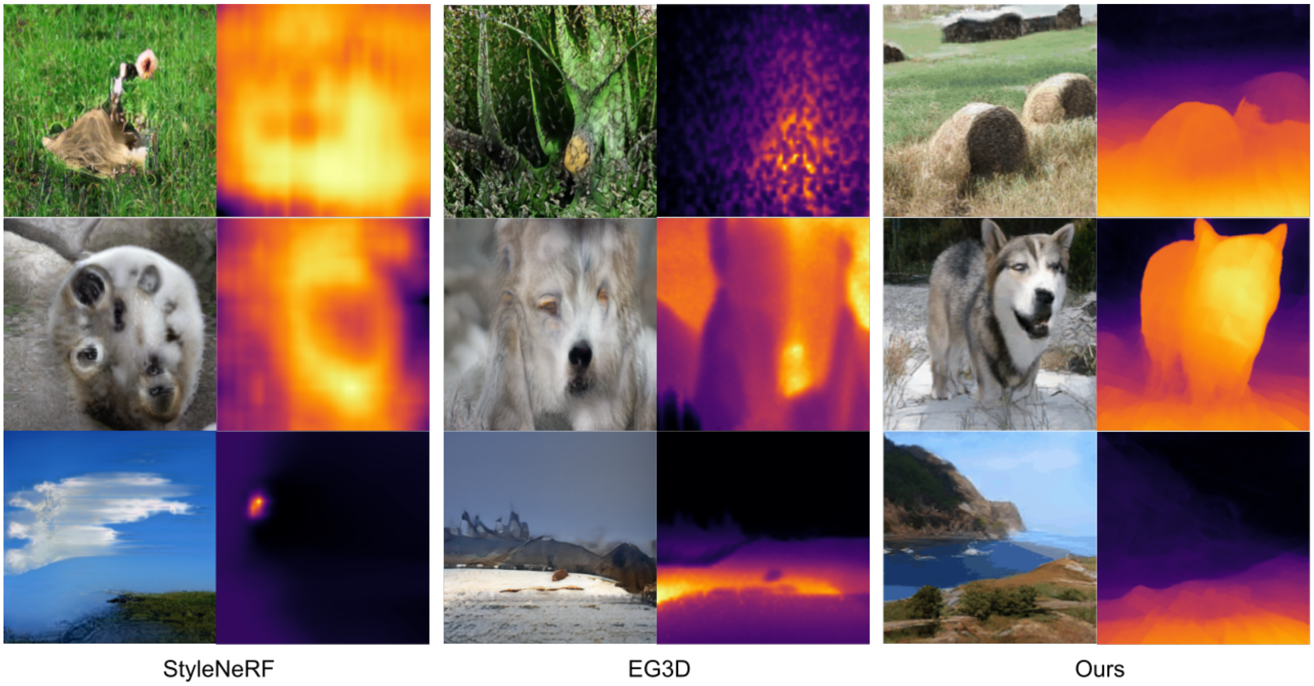

Figure 4: Generated samples and inverse depth from models trained on ImageNet. Ours model generates high-quality images and geometry. We generate samples from our method and then search for nearest neighbors in CLIP[27] space among generated samples of the baseline methods for more consistent evaluation of the quality improvement.

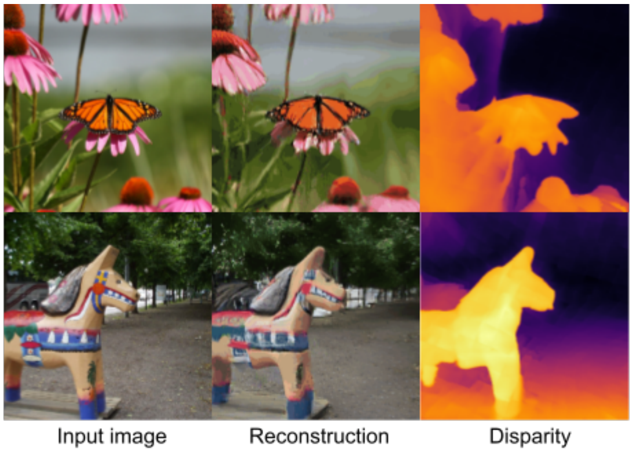

Figure 5: Reconstructions and estimated inverse depth on single images by our conditional NeRF-based autoencoder. Though our model is trained on ImageNet and achieves comparable performance on unseen ImageNet images, we show OpenImages results for licensing reasons.

methods without substantial changes.

Better geometry does not imply better FID. Additionally, learning geometry without a depth loss may be unreliable. For example, StyleNeRF [18] found learning of geometry was unreliable without training tricks such as progressive growing. During our ImageNet experiments, we also observed that the geometry StyleNeRF learns is sensitive to hyperparameters, and it often learns to produce flat depths. We were unable able to design a depth loss for StyleNeRF which improved the learned geometry. EG3D [6] showed that removing GT poses as input to the discriminator is enough to cause the geometry to degenerate to a flat plane.

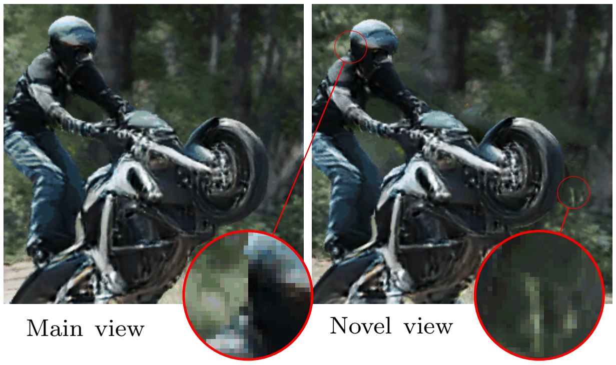

Figure 6: Example camera manipulations of a reconstructed scene. Our approach naturally handles sharp occlusions (left spyglass) and inpainting of disoccluded pixels (right spyglass) without supervision of novel views.

We conduct ablations on VQ3D in Table 3 starting from our main architecture (row 1). Using a CNN encoder and decoder like VQGAN [14] rather than ViT (row 2) is unstable and leads to codebook collapse. Eliminating the GAN loss (row 3) or depth scale loss (row 4) leads to a higher learned inverse depth scale and thus perceptually flat novel

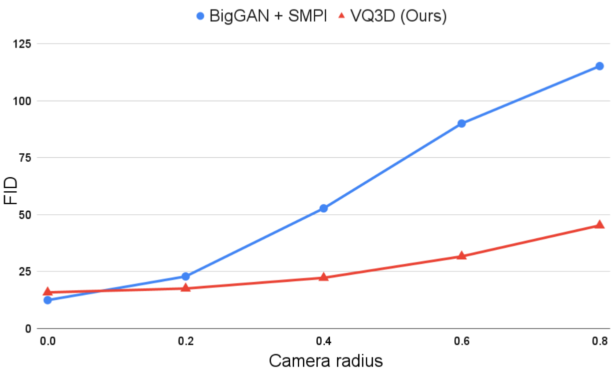

Figure 7: FID of our model versus a 2D generator baseline, BigGAN [3], plus a novel view synthesis model, Single-image MPI [44]. We see that the FID of our novel views degrades more gracefully as the sampling radius is increased.

Generation	**FID ↓**	**Depth accuracy ↓**
pi-GAN	101.4	1.41
GIRAFFE	132.1	1.78
EG3D	82.2	1.93
VQ3D (Ours)	31.7	1.90
pi-GAN + depth loss	97.8	0.88
GIRAFFE + depth loss	132.0	1.16
EG3D + depth loss	91.8	0.88
VQ3D (Ours) + depth loss	35.4	0.16

Table 2: Evaluation of depth losses on ImageNet. While adding depth losses can improve the quality of geometry, it will not improve FID enough to close the gap between our method and the baselines. We were unable to design a depth loss which prevented flat depths for StyleNeRF.

views. Additionally, removing the GAN loss (row 3) leads to artifacts in inpainting dis-occluded pixels during camera motion and thus to worse Stage 2 FID. Eliminating the NeRF loss (row 5) leads to worse depth accuracy and a very high inverse depth scale. Eliminating the depth loss (row 6) improves FID, but causes the depths to collapse to a flat plane and leads to worse depth accuracy. A fully implicit representation instead of triplanes (row 7) gives very poor FID because we can only use a very small MLP due to the expense of volume rendering at 256×256 resolution.

We additionally compare against a 2D GAN baseline BigGAN [3] equipped with a novel view synthesis model, Single-image MPI [44], in Figure 7. As the sampling radius is increased, this hybrid model performs worse relative to VQ3D. While existing 2D generative models [3, 35, 50] have excellent FID, it is nontrivial to make them 3D-aware.

4.3. Other 3D benchmark datasets

Two other prominent 3D-aware benchmark datasets are FFHQ [20] and CompCars [48]. Due to the ethical and legal issues associated with manipulation and generative modeling of faces, we do not study FFHQ. On CompCars, our model is competitive with the state of the art (Table 4). In order to study a dataset of intermediate complexity between the simple CompCars and highly diverse ImageNet, we introduce a new synthetic dataset, Multiview ShapeNet-1 (MVS-1), which we synthesize via Kubric [17]. MVS-1 is a variant of Multiview ShapeNet (MVS) introduced in [34] which consists of random ShapeNet [7] objects rendered against random HDRI backgrounds; the main difference between MVS and MVS-1 is that MVS-1 has exactly 1 salient object per image. We provide more details about MVS-1 in the supplementary materials and we will release the dataset upon acceptance. Our model is the best performing on MVS-1, although StyleNeRF performs much closer to VQ3D on this synthetic dataset than on ImageNet (Table 4).

Ablation	**FID-S1 ↓**	**FID-S2 ↓**	**Depth Acc. ↓**	**Disp. scale ↓**
(1) VQ3D (Ours)	11.2	35.4	0.18	1.00
(2) CNN enc, dec	(diverges)	-	-	
(3) W/o $\mathcal{L}_{\text{gan}}$	10.6	36.2	0.22	1.27
(4) W/o $\mathcal{L}_{\text{scale}}$	9.4	34.4	0.23	1.21
(5) W/o $\mathcal{L}_{\text{nerf}}$	9.2	36.6	0.28	4.88
(6) W/o $\mathcal{L}_{\text{depth}}$	4.0	33.0	1.91	0.61
(7) W/o Triplanes	274	275	1.00	2.15

Table 3: VQ3D ablation study. Removing components compromises the model capacity, 3D awareness, or novel view quality. FID-S1/2 is FID for Stage 1/2.

Model	piGAN	GIRAFFE	StyleNeRF	GIRAFFE HD	EG3D	VQ3D
CompCars	16.9	$26^{\dagger}$	$8^{\dagger}$	$7.2^{\dagger}$	32.2	7.3
MVS-1	104.4	54.7	11.5	-	42.6	9.8

Table 4: FID scores of 3D generative models on CompCars and MVS-1. † indicates numbers taken from the respective papers, we trained other models ourselves.

5. Discussion

Limitations and ethical considerations. While some benchmark methods [26, 18] have shown the ability to model 360-degree rotation of generated scenes when trained on specific single-class datasets like CompCars [48], our need to model 1000 object classes makes large viewpoint manipulation difficult. It is an interesting future direction to enable 360-degree rotation on general object classes. Furthermore, VQ3D training requires a depth estimator, like all ImageNet-focused concurrent works, i.e. IVID [19], 3DGP[41]. Finally, VQ3D training is multi-stage and relatively expensive.

Conclusion. We have presented VQ3D, a framework for 3D-aware representation learning and generation. VQ3D sets a state-of-the-art by a wide margin on the large and diverse ImageNet dataset, relative to existing strong geometry-aware baselines. We conduct extensive analysis and ablation verifying our contributions.

Acknowledgements

We would like to thank Peter Anderson, Jason Baldridge, Eric Ryan Chan, Alireza Fathi, Ben Mildenhall, and Guandao Yang for helpful discussions, Victor Gomes and Erica Moreira for help with resource coordination, and others in Google Research for support throughout this project.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 5, 6

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 1, 8

[4] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3990, 2022. 2

[5] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 1, 2, 6

[6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 1, 2, 6, 7

[7] 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. 8

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[10] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10673–10683, 2022. 2

[11] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 3

[12] 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. 2, 3, 5

[13] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. 3

[14] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2, 3, 7

[15] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 3, 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1, 2, 4

[17] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 8

[18] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1, 2, 6, 7, 8

[19] Binbin Huang Xin Tong Jianfeng Xiang, Jiaolong Yang. 3d-aware image generation using 2d diffusion models. 2023. 3, 8

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 8

[21] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9778–9787, 2019. 3, 4

[22] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokra, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5742–5752. PMLR, 18–24 Jul 2021. 3

[23] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. *arXiv preprint arXiv:2203.01941*, 2022. 3

[24] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 2, 5

[25] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. arXiv, 2019. 1, 2

[26] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 6, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 7

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 4

[31] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 4

[32] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look, 2022. 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 1

[34] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. *CVPR*, 2022. 8

[35] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. *CoRR*, abs/2202.00273, 2022. 1, 2, 6, 8

[36] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[37] Yichun Shi, Divyansh Aggarwal, and Anil K. Jain. Lifting 2d stylegan for 3d-aware face generation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. arXiv, 2020. 6

[38] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020. 2

[39] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2

[40] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[41] 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*, 2023. 3, 8

[42] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. Epigraf: Rethinking training of 3d gans. *arXiv preprint arXiv:2206.10535*, 2022. 2

[43] Ayush Tewari, Xingang Pan, Ohad Fried, Maneesh Agrawala, Christian Theobalt, et al. Disentangled3d: Learning a 3d generative model with disentangled geometry and appearance from monocular images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1516–1525, 2022. 2

[44] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[45] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 6

[47] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *CVPR*, 2022. 2

[48] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification, 2015. 1, 2, 8

[49] Hong-Xing Yu, Leonidas J Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. *arXiv preprint arXiv:2107.07905*, 2021. 3

[50] 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. *arXiv preprint arXiv:2110.04627*, 2021. 1, 2, 3, 4, 5, 6, 8

[51] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 1, 3