

# NeRDi: Single-View NeRF Synthesis with Language-Guided Diffusion as General Image Priors

Congyue Deng<sup>2\*</sup> Chiyu “Max” Jiang<sup>1</sup> Charles R. Qi<sup>1</sup> Xincheng Yan<sup>1</sup> Yin Zhou<sup>1</sup>  
Leonidas Guibas<sup>2,3</sup> Dragomir Anguelov<sup>1</sup>

<sup>1</sup>Waymo    <sup>2</sup>Stanford University    <sup>3</sup>Google Research

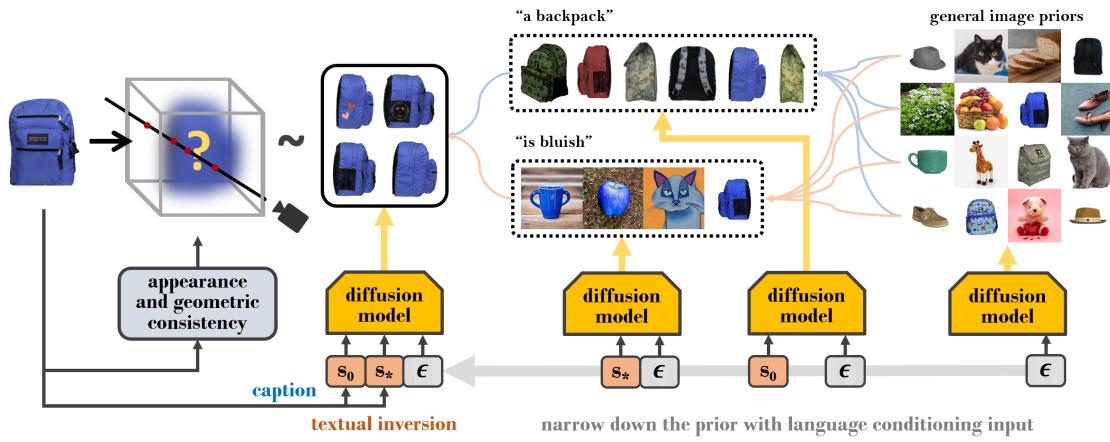


Figure 1. **From left to right:** We present a single-image NeRF synthesis framework for in-the-wild images without 3D supervision by leveraging general priors from large-scale image diffusion models. Given an input image, we optimize for a NeRF by minimizing an image distribution loss for arbitrary-view renderings with the diffusion model conditioned on the input image. We design a two-section semantic feature as the conditioning input to the diffusion model. The first section is the image caption  $s_0$  which carries the overall semantics; the second section is a text embedding  $s_*$  extracted from the input image with textual inversion, which captures additional visual cues. Our two-section semantic feature provides an appropriate image prior, allowing the synthesis of a realistic NeRF coherent to the input image.

## Abstract

2D-to-3D reconstruction is an ill-posed problem, yet humans are good at solving this problem due to their prior knowledge of the 3D world developed over years. Driven by this observation, we propose **NeRDi**, a single-view NeRF synthesis framework with general image priors from 2D diffusion models. Formulating single-view reconstruction as an image-conditioned 3D generation problem, we optimize the NeRF representations by minimizing a diffusion loss on its arbitrary view renderings with a pretrained image diffusion model under the input-view constraint. We leverage off-the-shelf vision-language models and introduce a two-section language guidance as conditioning inputs to the diffusion model. This is essentially helpful for improving multiview content coherence as it narrows down the general image prior conditioned on the semantic and visual features of the single-view input image. Additionally, we introduce a geometric loss based on estimated depth maps to regularize the underlying 3D geometry of the NeRF. Experimental

results on the DTU MVS dataset show that our method can synthesize novel views with higher quality even compared to existing methods trained on this dataset. We also demonstrate our generalizability in zero-shot NeRF synthesis for in-the-wild images.

## 1. Introduction

Novel view synthesis is a long-existing problem in computer vision and computer graphics. Recent progresses in neural rendering such as NeRFs [23] have made huge strides in novel view synthesis. Given a set of multi-view images with known camera poses, NeRFs represent a static 3D scene as a radiance field parametrized by a neural network, which enables rendering at novel views with the learned network. A line of work has been focusing on reducing the required inputs to NeRF reconstructions, ranging from dense inputs with calibrated camera poses to sparse images [12, 26, 52] with noisy or without camera

\*Work done as an intern at Waymo.

poses [48]. Yet the problem of NeRF synthesis from *one single view* remains challenging due to its ill-posed nature, as the one-to-one correspondence from a 2D image to a 3D scene does not exist. Most existing works formulate this as a reconstruction problem and tackle it by training a network to predict the NeRF parameters from the input image [9, 52]. But they require matched multiview images with calibrated camera poses as supervision, which is inaccessible in many cases such as images from the Internet or captured by non-expert users with mobile devices. Recent attempts have been focused on relaxing this constraint by using unsupervised training with novel-view adversarial losses and self-consistency [22, 51]. But they still require the test cases to follow the training distribution which limits their generalizability. There is also work [45] that aggregates priors learned on synthetic multi-view datasets and transfers them to in-the-wild images using data distillation. But they are missing fine details with poor generalizability to unseen categories.

Despite the difficulty of 2D-to-3D mapping for computers, it is actually not a difficult task for human beings. Humans gain knowledge of the 3D world through daily observations and form a common sense of how things should look like and should not look like. Given a specific image, they can quickly narrow down their prior knowledge to the visual input. This makes humans good at solving ill-posed perception problems like single-view 3D reconstruction. Inspired by this, we propose a single-image NeRF synthesis framework without 3D supervision by leveraging large-scale diffusion-based 2D image generation model (Figure 1). Given an input image, we optimize for a NeRF by minimizing an image distribution loss for arbitrary-view renderings with the diffusion model conditioned on the input image. An unconstrained image diffusion is the ‘general prior’ which is inclusive but also vague. To narrow down the prior knowledge and relate it to the input image, we design a two-section semantic feature as the conditioning input to the diffusion model. The first section is the image caption which carries the overall semantics; the second is a text embedding extracted from the input image with textual inversion [10], which captures additional visual cues. These two sections of language guidance facilitate our realistic NeRF synthesis with semantic and visual coherence between different views. In addition, we introduce a geometric loss based on the estimated depth of the input view for regularizing the underlying 3D structure. Learned with all the guidance and constraints, our model is able to leverage the general image prior and perform zero-shot NeRF synthesis on single image inputs. Experimental results show that we can generate high quality novel views from diverse in-the-wild images. To summarize, our key contributions are:

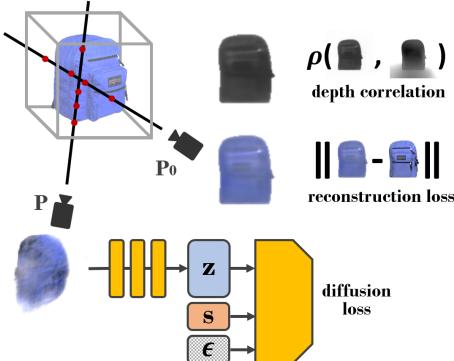
- We formulate single-view reconstruction as a conditioned 3D generation problem and propose a single-image NeRF

synthesis framework without 3D supervision, using 2D priors from diffusion models trained on large image datasets.

- We design a two-section semantic guidance to narrow down the general prior knowledge conditioned on the input image, enforcing synthesized novel views to be semantically and visually coherent.
- We introduce a geometric regularization term on estimated depth maps with 3D uncertainties.
- We validate our zero-shot novel view synthesis results on the DTU MVS [13] dataset, achieving higher quality than supervised baselines. We also demonstrate our capability of generating novel-view renderings with high visual quality on in-the-wild images.

## 2. Related Work

**Novel view synthesis with NeRF.** The recently proliferating NeRF representation [23] has shown great success in novel view synthesis, which is a long-existing task in computer graphics and vision. Combining differentiable rendering [16, 53, 54, 55] with neural network scene parametrizations, NeRF is able to recover the underlying 3D scene from a collection of posed images and render it at novel views realistically. A number of follow-up works have been focusing on relaxing NeRF inputs to less informative data such as unposed images [21, 48, 50] or sparse views [7, 12, 26, 34]. As less data gives rise to a more complex optimization landscape, a variety of regularization losses have been studied, for example: RegNeRF [26] regularizes the geometry and appearance of patches, DDP [34] and DS-NeRF [7] regularize the depth maps, DietNeRF [12] enforces semantic consistency between views by minimizing a CLIP [30] feature loss, and GNeRF [21] adopts a patch-based adversarial loss. Another line of work learns NeRF-based novel-view prediction for few- or single-image inputs by pre-training a scene prior on a large dataset of 3D scenes containing dense views [5, 6, 18, 44, 46, 52]. With additional self-supervision techniques such as equivariance [9] or cycle-consistency [22], the learning of scene priors can be done simply from sparse- or single-view data, or even purely from unposed image collections with an image adversarial loss [2, 3, 27, 39]. These two lines of works both have their specialties and constraints: the first is generalizable to any scene configurations, but is also less competitive in the more challenging scenarios such as single-image novel view synthesis with high quality requirements; the second, on the other hand, has strong ability of inferring unseen novel views from very limited inputs, but is also restricted to certain scene categories modeled by their scene priors learned from the training data. In our work, we leverage a diffusion-based image prior for NeRF synthesis that is general enough for modeling variations of in-the-wild images while having the adaptivity to each specific input image.



**Figure 2. Method overview.** We represent the underlying 3D scene as a NeRF and optimize for its parameters with three losses: a reconstruction loss at the fixed input view; a diffusion loss at arbitrarily sampled views which also takes a conditioning text input generated from the input image with our two-section feature extraction; and finally, a depth correlation loss at the input view regularizing the 3D geometry.

**Diffusion-based generative models.** Denoising diffusion probabilistic models [11, 41], or score-based generative models [42, 43], have recently caught a surge of interests due to their simple designs and excellent performances across a variety of computer vision tasks such as image generation [11, 41, 42, 43], completion [36, 43], and editing [14, 20]. In visual content creation, language-guided image diffusion models such as DALL-E2 [32], Imagen [37] and Stable Diffusion [35] have shown great success in generating photo realistic images with strong semantic correlation to the given text-prompt inputs. In addition to the success of 2D image diffusion models, more recent works have also extend diffusion models to 3D content generation. [19, 57] generate 3D pointclouds with point diffusions. 3DiM [49] shows uncertainty-aware novel view synthesis with image diffusions conditioned on input views and poses, but it does not have guaranteed multiview consistency as no underlying 3D representation is adopted. More related to ours are DreamFusion [28] and GAUDI [1] that also generate NeRFs with diffusions: [28] generates NeRFs under language guidance by optimizing for their renderings at randomly sampled views with a 2D image diffusion model [37]; [1] trains a diffusion model on the latent space of NeRF scenes, but the learned scene distribution is limited to a set of indoor 3D scenes and does not generalize to in-the-wild images. Similar to [28], we also leverage 2D image diffusions to optimize for the NeRF renderings at novel views, but instead of unconstrained NeRF generation with user-specified language inputs, we study how to faithfully capture the features of single-view image inputs and use it to constrain the novel-view image distributions.

### 3. Method

An overview of our method is shown in Figure 2. Given an input image  $\mathbf{x}_0$ , we would like to learn a NeRF repre-

sentation  $F_\omega : (x, y, z) \rightarrow (\mathbf{c}, \sigma)$  as its 3D reconstruction<sup>†</sup>. The NeRF holds the rendering equation that, for any camera view with pose  $\mathbf{P}$ , one can sample camera rays  $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$  and render the image  $\mathbf{x}$  at this view with

$$\hat{\mathbf{C}}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(t)\mathbf{c}(t)dt \quad (1)$$

where  $T(t) = \exp\left(-\int_{t_n}^t \sigma(s)ds\right)$ . For more details, please refer to Mildenhall *et al.* [23]. For simplicity, we denote this whole rendering equation by  $\mathbf{x} = f(\mathbf{P}, \omega)$  which means NeRF  $f$  renders image  $\mathbf{x}$  at camera pose  $\mathbf{P}$  with parameters  $\omega$ . Instead of predicting the NeRF parameters  $\omega$  from  $\mathbf{x}_0$  in a forward pass, we formulate this as a conditioned 3D generation problem

$$f(\cdot, \omega) \sim \text{3D scene distribution} \mid f(\mathbf{P}_0, \omega) = \mathbf{x}_0 \quad (2)$$

where we optimize the NeRF to follow a 3D scene distribution conditioned on that its rendering  $f(\mathbf{P}_0, \omega)$  at a given view  $\mathbf{P}_0$  should be the input image  $\mathbf{x}_0$

Directly learning the 3D scene distribution prior requires large 3D datasets, which is less straightforward to acquire and restricts its application to unseen scene categories. To enable better generalizability to in-the-wild scenarios, we instead leverage 2D image priors and reformulate the objective into

$$\forall \mathbf{P}, f(\mathbf{P}, \omega) \sim \mathbb{P} \mid f(\mathbf{P}_0, \omega) = \mathbf{x}_0 \quad (3)$$

where the optimization is conducted on images  $f(\mathbf{P}, \omega)$  rendered at arbitrarily sampled views, pushing them to follow an image prior  $\mathbb{P}$  while satisfying the constraint  $\mathbf{x}_0 = f(\mathbf{P}_0, \omega)$ . The overall objective can be written as maximizing the conditional probability

$$\max_{\omega} \mathbb{E}_{\mathbf{P}} \mathbb{P}(f(\mathbf{P}, \omega) \mid f(\mathbf{P}_0, \omega) = \mathbf{x}_0, \mathbf{s}). \quad (4)$$

Here,  $\mathbf{s}$  is an additional semantic guidance term that we apply to further restrict the prior image distribution to fit the generation context. In contrast to DreamFusion [28] which also utilizes language-guided image diffusion model as 2D image priors for sampled views, our main contribution stands in our approach for further constraining the identity of the generated 3D volume to be consistent with the inputs.

We cover more details on this novel-view distribution loss in Sec. 3.1. We utilize natural language descriptions of the scene as the semantic guidance  $\mathbf{s}$ . More details on this will be discussed in Sec. 3.2. In addition, as the image diffusion model only operates on the rendered rgb colors, we further apply a geometric regularization with a depth map estimated at the input view to facilitate the NeRF optimization (Sec. 3.3).

### 3.1. Novel View Distribution Loss

Denoising Diffusion Probabilistic Models (DDPM) are a type of generative models that learn a distribution over

<sup>†</sup>Here we use a Lambertian NeRF without view direction inputs for enforcing stronger multiview consistency.

training data samples. Recently, there are many advances in language guided image synthesis with diffusion models. We build our method upon the recent Latent Diffusion Model (LDM) [35] for its high quality and efficiency in image generation. It adopts a pre-trained image auto-encoder with an encoder  $\mathcal{E}(\mathbf{x}) = \mathbf{z}$  mapping images  $\mathbf{x}$  into latent codes  $\mathbf{s}$  and a decoder  $\mathcal{D}(\mathcal{E}(\mathbf{x})) = \mathbf{x}$  recovering the images. The diffusion process is then trained in the latent space by minimizing the objective

$$\mathbb{E}_{\mathbf{z} \sim \mathcal{E}(\mathbf{x}), \mathbf{s}, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c_\theta(\mathbf{s}))\|_2^2]. \quad (5)$$

where  $t$  is a diffusion time scale,  $\epsilon \sim \mathcal{N}(0, 1)$  is a random noise sample,  $\mathbf{z}_t$  is the latent code  $\mathbf{z}$  noised to time  $t$  with  $\epsilon$ , and  $\epsilon_\theta$  is the denoising network with parameters  $\theta$  to regress the noise  $\epsilon$ . The diffusion model also takes a conditioning input  $\mathbf{s}$  which is encoded as  $c_\theta(\mathbf{s})$  and serves as guidance in the denoising process. For text-to-image generation models such as the LDM,  $c_\theta$  is a pre-trained large language model that encodes the conditional text  $\mathbf{s}$ .

In a pre-trained diffusion model, the network parameters  $\theta$  are fixed, and we can instead optimize for the input image  $\mathbf{x}$  with the same objective which transforms  $\mathbf{x}$  to follow the image distribution priors conditioned on  $\mathbf{s}$ . Let  $\mathbf{x} = f(\mathbf{P}, \omega)$  be our NeRF rendering at arbitrarily sampled view  $\mathbf{P}$ , we can back propagate gradients to the NeRF parameters  $\omega$  and thus get a stochastic gradient descent on  $\omega$ .

### 3.2. Semantics-Conditioned Image Priors

We argue that the prior distribution over all in-the-wild images is not specific enough to guide the novel view synthesis from an arbitrary image. We thus introduce a well-designed guidance  $\mathbf{s}$  that narrows down the generic prior over natural images to a prior of images related to the input image  $\mathbf{x}_0$ . Here we choose text as the guidance, which is flexible for describing arbitrary input images. Text-to-image diffusion models such as LDM utilize a pre-trained large language model as the language encoder to learn a conditional distribution over images conditioned on language. This serves as a natural gateway for us to utilize language as a means to restrict the image prior space.

The most straightforward way of getting a text prompt from the input image is to use an image captioning or classification network  $\mathcal{S}$  trained on (image, text) datasets and predict a text  $\mathbf{s}_0 = \mathcal{S}(\mathbf{x}_0)$ . However, while text description can summarize the semantics of the image, it leaves a huge space of ambiguities, making it hard to include all the visual details in the image especially with limited prompt length. In Figure 3 top row, we show the images generated with the caption “a collection of products” from the input image on the left. While their semantics are highly accurate with respect to the language description, the generated images have very high variances in their visual patterns and low correlations to the input image.

Textual inversion [10], on the other hand, optimizes for the text embedding of one or few images from a text-based

image diffusion model. With the LDM Equation 5, we can optimize for the text embedding  $\mathbf{s}_*$  for the input image  $\mathbf{x}_0$  by

$$\mathbf{s}_* = \arg \min_{\mathbf{s}} \mathbb{E}_{\mathbf{z} \sim \mathcal{E}(\mathbf{x}_0), \mathbf{s}, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c_\theta(\mathbf{s}))\|_2^2] \quad (6)$$

In Figure 3 middle row, images generated with textual inversion are shown. The colors and visual cues of the input image are well captured (orange-colored elements, food, and even the brand logos). However, the semantics at the macro level is sometimes wrong (second column is a person playing sports). One reason is that, different from the multi-image scenarios where textual inversion can discover the common contents of these images, it is unclear for one single image what the key features are that the text embedding should focus on.

To reflect both semantic and visual characteristics of the input image in the novel view synthesis task, we combine these two methods by concatenating their text embeddings to form a joint feature  $\mathbf{s} = [\mathbf{s}_0, \mathbf{s}_*]$  and use it as the guidance in the diffusion process in Equation 5. Figure 3 bottom row shows the images generated with this joint feature, with balanced semantics and visual cues.

### 3.3. Geometric Regularization

While image diffusion shapes the appearance of the NeRF, multiview consistency is difficult to enforce as the underlying 3D geometry can be different even with the same image rendering [15, 24], making the gradient back-propagation (from the image diffusion to the NeRF parameters  $\omega$ ) highly non-controllable. To this end, we further incorporate a geometric regularization term on the input view depth to alleviate this issue. We adopt the Dense Prediction Transformer (DPT) model [33] trained on 1.4 million images for zero-shot monocular depth estimation and apply it to the input image  $\mathbf{x}_0$  to estimate a depth map  $\mathbf{d}_{0,\text{est}}$ . We use this estimated depth to regularize the depth

$$\hat{\mathbf{d}}_0 = \int_{t_n}^{t_f} \sigma(t) dt. \quad (7)$$

rendered by the NeRF at input view  $\mathbf{P}_0$ . Due to the ambiguities of the estimated depth (including scales, shifts, camera intrinsics) and estimation error (Figure 4), we cannot back project pixels with depth to 3D and compute the regularization directly. Instead, we maximize the *Pearson* correlation between the estimated depth map and the NeRF-rendered depth

$$\rho(\hat{\mathbf{d}}_0, \mathbf{d}_{0,\text{est}}) = \frac{\text{Cov}(\hat{\mathbf{d}}_0, \mathbf{d}_{0,\text{est}})}{\sqrt{\text{Var}(\hat{\mathbf{d}}_0)\text{Var}(\mathbf{d}_{0,\text{est}})}} \quad (8)$$

which measures if the rendered depth distribution and the noisy estimated depth distribution are linearly correlated.

## 4. Experiments

Now we demonstrate our efficacy in synthesizing realistic NeRFs with single-view inputs. Section 4.1 presents a



Figure 3. **Image generation with different semantic guidance.** **Top row:** Images generated with caption “**a collection of products**”. The images follows the semantics well, but their content are of very high variance (can be any kind of products). **Middle row:** Images generated purely with the latent embedding from **textual inversion**. The color distribution and visual cues of the input image are well captured, but the semantics is not preserved (second column, the image is a person playing sports). **Bottom row:** Images generated with combined image caption and textual inversion. Both semantic and visual features of the input image are addressed.

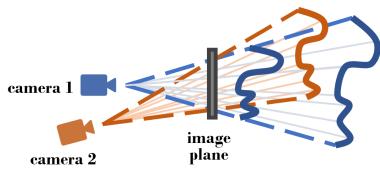


Figure 4. **Ambiguity in estimated depth map.**

Table 1. Single-image novel view synthesis results on DTU.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
NeRF	8.000	0.286	0.703
pixelNeRF	15.550	0.537	0.535
pixelNeRF, $\mathcal{L}_{\text{MSE}}$ ft	<b>16.048</b>	<b>0.564</b>	0.515
DietPixelNeRF	14.242	0.481	0.487
Ours	14.472	0.465	<b>0.421</b>

quantitative comparison between our method and the state-of-the-art single-view NeRF reconstruction methods on a synthetic dataset. Section 4.2 shows a qualitative comparison as well as more synthesis results of our method on in-the-wild images.

#### 4.1. Synthetic Scenes

**Setup.** We evaluate our method on the DTU MVS dataset [13] with 15 test scenes as specified in [52]. For each input image, we use GPT-2 [31] to generate a caption. We manually correct the obvious mistakes made by GPT-2 while trying our best to avoid introducing additional details. The scenes and their captions are listed in the [supplementary material](#).

**Implementation details.** For the NeRF model, we implement the multi-resolution grid sampler as described in [25]. For the diffusion model, we employ the text-guided diffusion model from [35] which was pre-trained on the LAION-400M dataset [38]. While [35] operates on  $512 \times 512$  images, NeRF’s volumetric rendering at this resolution would incur an extensive computational burden. Thus, at the ran-

domly sampled novel views, we render  $128 \times 128$  images and resize them to  $512 \times 512$  before feeding them to the encoder of [35]. At the input view, we render at the same resolution as the input image to compute the image reconstruction and depth correlation losses.

**Baselines.** We compare with two state-of-the-art single-view NeRF reconstruction algorithms, PixelNeRF [52] and its fine-tuned model with CLIP [30] feature consistency loss as proposed by DietNeRF [12], both of which trained on the training set data from the DTU MVS dataset. To gain better convergence, we use the predictions from [52] as an initialization for our 3D scene optimization. But our method is directly applied to the test scenes without any additional fine-tuning on the DTU training set.

**Results.** Table 1 shows the quantitative comparison between our method and the baselines. Following the convention, we report the standard image quality metrics PSNR and SSIM [47]. Our PSNR and SSIM are slightly lower than pixelNeRF [52] which directly learns the scene distributions from the DTU training set and are on par with DietPixelNeRF [12] which enforces semantic consistency between views. However, we emphasize that these two metrics are less indicative in our scenario as they are local pixel-aligned similarity metrics between the synthesized novel views and the ground truth images but uncertainties naturally exist in single-view 3D inference. The middle column of the first scene in Figure 5 shows an example of such uncertainty. The height of the tallest snack bag in the input image cannot be inferred as its top extrudes beyond the camera view. The width of the toy pig in the left column of the third scene is another example which cannot be inferred from the input side view. In both cases our method guesses its novel view (bottom row) in a reasonable sense but different from the ground truth (top row). In addition, we also measure novel views with LPIPS [56], which is a perceptual

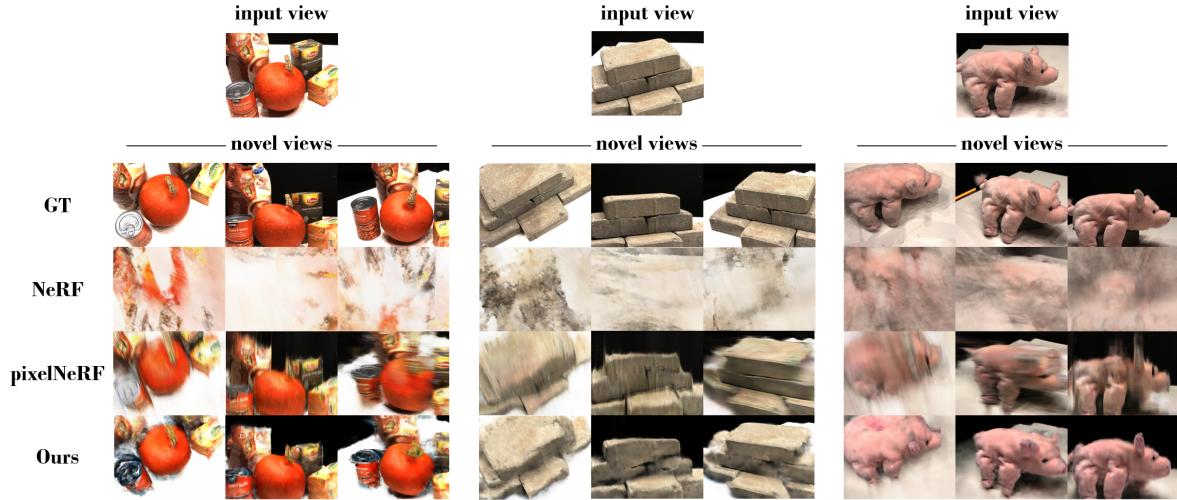


Figure 5. **Single-image novel view synthesis results on the DTU test scenes.** Vanilla NeRF cannot recover scenes from single image inputs due to the ill-posed nature of this problem. While pixelNeRF can infer the novel view images with the prior from the DTU training set of similar scenes, its synthesized renderings remain noisy and blurry. With a pixelNeRF initialization, our method is able to synthesize cleaner novel views with realistic geometries and appearances, despite having never been trained on this dataset. **Uncertainties in novel view inference:** (The first scene, middle column) the exact height of the tallest snack bag cannot be inferred as the top goes outside of the camera view. (The third scene, left column) the width of the toy pig from the top view is undecidable from the input view. In both cases, our method guesses a reasonable answer in the synthesized novel view that is different from the ground truth.

metric computing the Mean Squared Error (MSE) between normalized features from all layers of a pre-trained VGG encoder [40]. Our method shows a significant improvement on this metric compared to the baselines as the diffusion model helps to improve image qualities while the language guidance maintains the multi-view semantic consistency.

Figure 5 shows a qualitative comparison between our method and the baselines. With the scene initialization from [52], our method removes the noises and blurriness, synthesizing high quality novel views.

## 4.2. Images in the Wild

**Qualitative comparisons.** Figure 6 shows a qualitative comparison between our method and existing state-of-the-art single-image to 3D synthesis methods for in-the-wild images [12, 45]. Input images are adopted from the Google Scanned Objects dataset [8] with their category labels ('bag' and 'hat') as captions. Similar to ours, Diet-NeRF [12] uses an input-view constrained NeRF optimization technique where they minimize the CLIP [30] feature between arbitrary view renderings. While CLIP features enforce consistent appearances, they fail to capture the global semantics of the object. SS3D [45] is a forward-prediction model for 3D geometries that transfers the priors learned on synthetic datasets to in-the-wild images with knowledge distillation. While it generates more structured global geometries, it fails to capture the fine geometric details of the input image. The geometries of the hats in the bottom rows are also incorrect, with only the silhouette shape preserved but the structure of 'hat' shape missing.

**More results.** Figure 7a shows our results on images of objects from the internet. The text prompts are words or phrases used to search for the images. The backgrounds are masked out using an off-the-shelf dichotomous image segmentation network from [29]. For each input, we show 3 different novel views that are distant from the input view. Figure 7b shows our results on images with more complex contents and backgrounds from the COCO dataset [17] which contains (image, caption) pairs. Within camera views close to the input, our model is still able to generate realistic renderings. But it can hardly generalize to distant views due to the limited capacity of the NeRF scene box.

## 4.3. Ablation Studies

We conduct ablation studies to show the efficacy of our two-section semantic guidance and geometric regularization.

**Semantic guidance.** Figure 8a shows the ablation of the two text embeddings  $s_0$  from image captions and  $s_*$  from textual inversion. Without the captions  $s_0$ , the model fails to learn the overall semantics and cannot generate a meaningful object. While both the full model and the caption-only one (without textual inversion) successfully generate backack novel views, the results without textual inversion  $s_*$  have more blurriness and noises. A zoom-in comparison is shown in Figure 8b.

Figure 8c shows another comparison of models with and without textual inversion  $s_*$  on the can example from Figure 7b left. In the object regions visible to the input view, the full model better recovers the fine details (the white letters

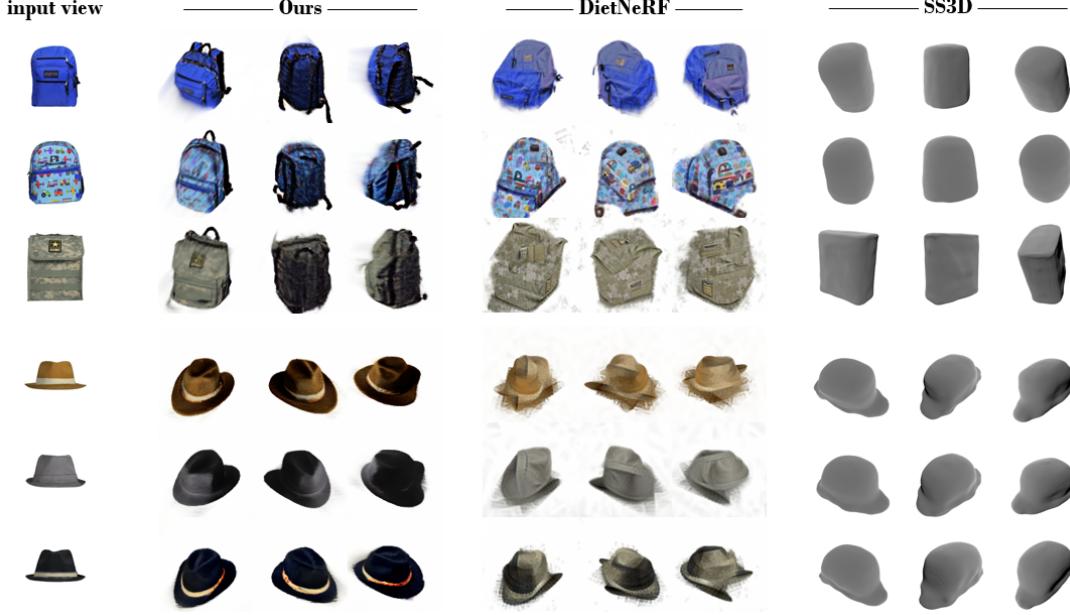


Figure 6. **Novel view synthesis results on objects from the Google Scanned Objects Dataset.** **Left:** Our results generated from single-word text inputs ‘backpack’ (top 3 rows) and ‘hat’ (bottom 3 rows). **Middle:** DietNeRF [12] minimizes the CLIP feature distances between the input view and arbitrarily sampled views. This results in novel view renders with consistent textures and styles, but fails to capture the global semantic meaning. *For a fair comparison, DietNeRF is also optimized with depth regularization.* **Right:** SS3D [45] predicts coarse geometries in a consistency manner, but it fails to recover all the fine geometric details. Additionally, the geometries of the hats in the bottom rows are incorrect, with only the silhouette shape preserved but the structure of ‘hat’ shape missing.



(a) **Results on object-centric images from the internet with single-word or short phrase captions.** Input backgrounds are removed with [29].

Figure 7. **Results on images in the wild.**

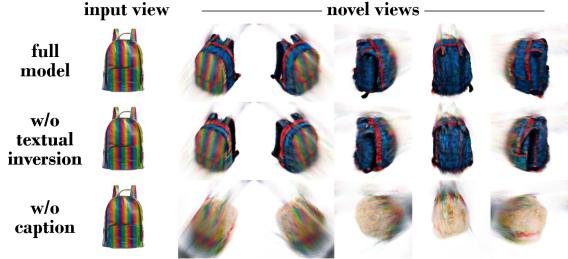
on the lateral); and in the invisible regions, the full model completes the appearances with coherent styles of the input (red and white textures at the back of the can), while the model without textual inversion does not have such appearance coherency. The model with textual inversion can even synthesize the pull tab at the top (second column of the zoom-in views) by inferring from the input side view that this is a can containing drinks.

**Geometric regularization.** Figure 9 shows an ablation on the geometric regularization term. Both image renderings and depth maps are visualized. The full model is able to synthesize realistic novel views with coherent 3D geome-

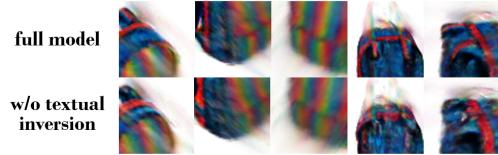
try. The model without the regularization on the input view depth can still generate realistic appearances at novel views with the diffusion model, but the underlying 3D geometry is erroneous and multi-view consistency is not enforced. As a sanity check, we also visualize the results with only the depth loss but without the diffusion model. The model is unable to generate a realistic NeRF due to the 3D ambiguities of monocular depth as stated in Section 3.3.

## 5. Conclusions

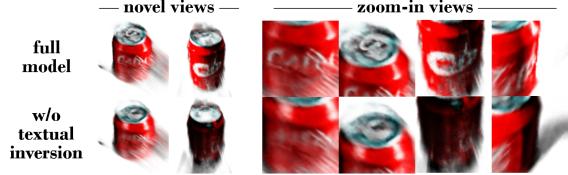
In this paper, we propose a novel framework for zero-shot single-view NeRF synthesis for images in the wild



(a) **Top row:** Full model. **Middle row:** Caption-only guidance without textual inversion. The model is still able to generate a shape strictly following the semantics and the input view appearance and geometric constraints, but struggles more in synthesizing the details. A zoom-in comparison is shown in 8b below. **Bottom row:** Textual-inversion-only without caption. Textual inversion fails to capture the global semantics.



(b) **A zoom-in comparison between full model results and results without textual inversion.** The full model shows better capability of synthesizing less blurry details.



(c) **Another comparison between models with and without textual inversion.** The input is from 7b left, bottom row. The full model is able to synthesize better texture details at visible regions as well as completing the invisible regions with similar textures, while the caption-only model renderings are more blurry and cannot fill in the invisible regions.

Figure 8. **Ablations on the two-section semantic guidance.**

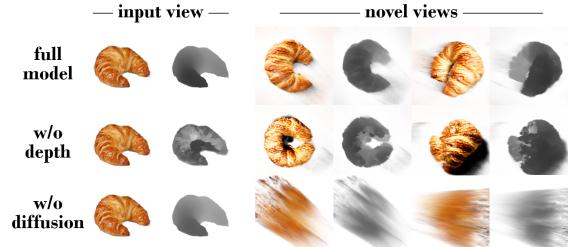
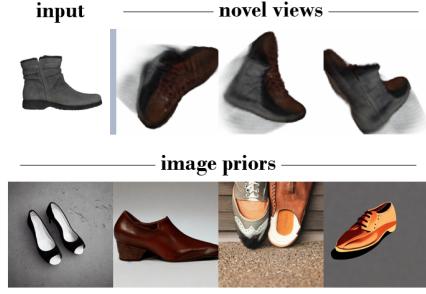


Figure 9. **Ablations on the geometric regularization.** Visualization of input view reconstruction and novel views on rendered images and depth maps. **Top row:** The full model is able to synthesize realistic novel views while preserving geometric coherency. **Middle row:** Without the depth correlation loss, the diffusion model is still able to generate reasonable appearances, but the underlying 3D geometry is erroneous and the novel views are inconsistent to the input view. **Bottom row:** The input-view depth estimation cannot guide novel view synthesis by itself without the diffusion model due to 3D ambiguities.

without 3D supervision. We leverage the general image priors in 2D diffusion models and apply them to the 3D NeRF



(a) **A failure case due to the biases in the image diffusion model.** **Top:** Novel view synthesis results with text prompt ‘a shoe in the style of <input>’. **Bottom:** Images generated by [35] with text prompt “a single shoe”. Yet half of the images have two shoes in it.



(b) **A failure case on a highly deformable instance.** While the overall body shape of the cat is captured, the synthesized cat has two heads and two tails.

Figure 10. **Failure cases.**

generation conditioned on the input image. To efficiently use these priors in synthesizing consistent views, we design a two-section language guidance as conditioning inputs to the diffusion model which unifies the semantic and visual features of the input image. To our knowledge, we are the first to combine semantic and visual features in the text embedding space and apply it to novel view synthesis. In addition, we introduce a geometric regularization term while addressing the 3D ambiguity of monocular-estimated depth maps. Our experimental results show that, with well-designed guidance and constraints, one can leverage general image priors to specific image-to-3D, enabling us to build generalizable and adaptable reconstruction frameworks.

**Limitations and future work.** As our method relies on multiple large pre-trained image models [29, 31, 33, 35], any biases in these models will affect our synthesis results. Figure 10a shows an example where the image diffusion model [35] can generate two shoes even the text prompt is “a single shoe”, resulting in our synthesized NeRF showing the features of multiple shoes. Our method is also less robust to highly deformable instances, as our language guidance focuses on semantics and styles but lacks a global description of physical states and dynamics. Figure 10b shows such a failure case. Renderings from each independent view are visually plausible but represent different states of the same instances.

Besides, while formulation-wise the optimization is applicable to any scenes, it is more suitable for object-centric images as it takes the underlying assumption that the scene has exactly the same semantics from any view, which is not true for large scenes with complex configurations due to



Figure 11. Images generated by [35] with ‘a pumpkin’.

view changes and occlusions. The text embedding learned from textual inversion is of the dimension of a single-world embedding, limiting its expressiveness in representing the subtleties complex contents.

## A. Additional Results

Figure 12 shows our additional results and comparisons for images in the wild. The results are presented in 4 groups, each group containing 3 objects from similar classes but with different content details and appearances. We use this to test the capability of each method in capturing the overall semantics and visual feature variations from input images.

**Comparison to DietNeRF [12].** For a fair comparison, DietNeRF is also optimized with the estimated depth map from the input image. While DietNeRF is able to maintain appearance consistency between different views, it fails to capture the overall geometry of the objects, especially when the object has complex geometric structures (such as the chairs in the 1st group, and the baskets in the 3rd group). In the 4th group (the skirts), our generated textures form the unseen back regions are also closer to the input image than DietNeRF.

Our method also addresses the naturally existing ambiguity in novel-view inference, especially for the occluded regions in the input view. For example, in the 3rd group in Figure 12, the unseen spaces of the baskets are filled with different fruits/flowers/vegetables, instead of duplicating the input views as DietNeRF [12]. As a feature or as an inductive bias, such synthesis results are also affected by the 2D distribution from the image diffusion model. For example, Figure 11 shows the image generation results by [35] with text prompt ‘a pumpkin’. Half of them are Jack-o’-lanterns. This makes our synthesized pumpkin also having the Jack-o’-lantern face at its back (the 3rd row of the 2nd group).

**Comparison to SS3D [45].** As a geometry-based method, SS3D captures better global geometries than DietNeRF even without the depth regularization, especially on the object classes covered by ShapeNet [4] where the

## References

- [1] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*, 2022. 3
- [2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2
- [3] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2
- [4] 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. 9
- [5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2
- [6] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 2
- [7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2
- [8] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arXiv preprint arXiv:2204.11918*, 2022. 6
- [9] Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, pages 2761–2770. PMLR, 2020. 2
- [10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 4
- [11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3
- [12] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 1, 2, 5, 6, 7, 9
- [13] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 2, 5
- [14] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–

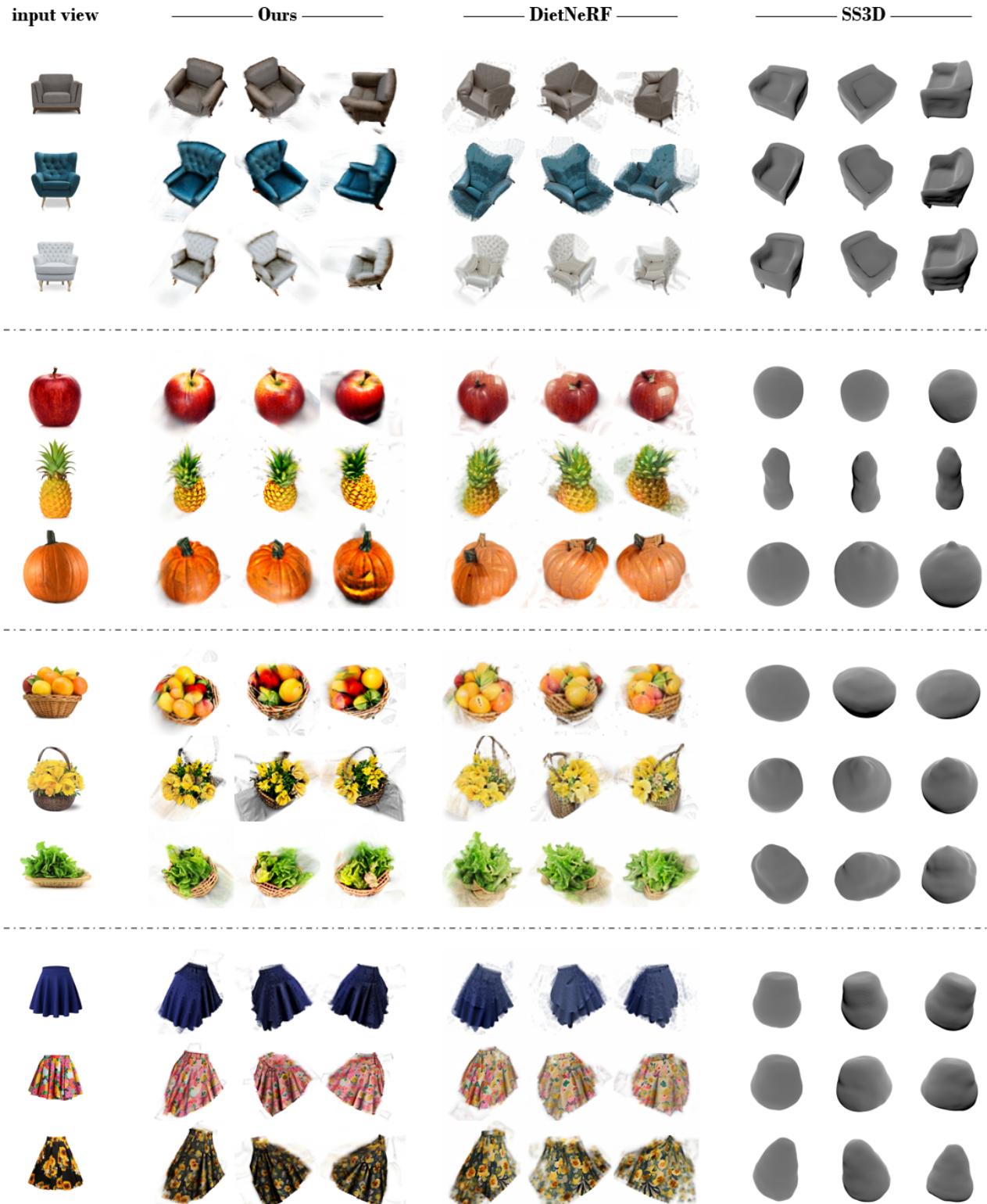


Figure 12. Additional results for images in the wild.

- 2435, 2022. 3
- [15] Steven M Lehar. *The world in your head: A gestalt view of the mechanism of conscious experience*. Psychology Press, 2003. 4
- [16] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 2
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 7
- [18] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 2
- [19] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 3
- [20] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3
- [21] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 2
- [22] Lu Mi, Abhijit Kundu, David Ross, Frank Dellaert, Noah Snavely, and Alireza Fathi. im2nerf: Image to neural radiance field in the wild. *arXiv preprint arXiv:2209.04061*, 2022. 2
- [23] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3
- [24] Niloy J Mitra and Mark Pauly. Shadow art. *ACM Transactions on Graphics*, 28(CONF):156–1, 2009. 4
- [25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 5
- [26] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Reg-nerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1, 2
- [27] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2
- [28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [29] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022. 6, 7, 8
- [30] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6
- [31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 5, 8
- [32] 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. 3
- [33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 4, 8
- [34] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 3, 4, 5, 8, 9
- [36] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3
- [37] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3
- [38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5
- [39] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2
- [40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*

- arXiv:2010.02502*, 2020. 3
- [42] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3
- [44] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2
- [45] Kalyan Alwala Vasudev, Abhinav Gupta, and Shubham Tulsiani. Pre-train, self-train, distill: A simple recipe for supersizing 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6, 7, 9
- [46] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2
- [47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [48] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2
- [49] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3
- [50] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2
- [51] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021. 2
- [52] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 2, 5, 6
- [53] Cheng Zhang, Bailey Miller, Kan Yan, Ioannis Gkioulekas, and Shuang Zhao. Path-space differentiable rendering. *ACM transactions on graphics*, 39(4), 2020. 2
- [54] Cheng Zhang, Lifan Wu, Changxi Zheng, Ioannis Gkioulekas, Ravi Ramamoorthi, and Shuang Zhao. A differential theory of radiative transfer. *ACM Transactions on Graphics (TOG)*, 38(6):1–16, 2019. 2
- [55] Cheng Zhang, Zihan Yu, and Shuang Zhao. Path-space differentiable rendering of participating media. *ACM Transactions on Graphics (TOG)*, 40(4):1–15, 2021. 2
- [56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5
- [57] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 3