Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Matthew Tancik
UC Berkeley
tancik@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present DietNeRF, a 3D neural scene representation estimated from a few images. Neural Radiance Fields (NeRF) learn a continuous volumetric representation of a scene through multi-view consistency, and can be rendered from novel viewpoints by ray casting. While NeRF has an impressive ability to reconstruct geometry and fine details given many images, up to 100 for challenging 360° scenes, it often finds a degenerate solution to its image reconstruction objective when only a few input views are available. To improve few-shot quality, we propose DietNeRF. We introduce an auxiliary semantic consistency loss that encourages realistic renderings at novel poses. DietNeRF is trained on individual scenes to (1) correctly render given input views from the same pose, and (2) match high-level semantic attributes across different, random poses. Our semantic loss allows us to supervise DietNeRF from arbitrary poses. We extract these semantics using a pre-trained visual encoder such as CLIP, a Vision Transformer trained on hundreds of millions of diverse single-view, 2D photographs mined from the web with natural language supervision. In experiments, DietNeRF improves the perceptual quality of few-shot view synthesis when learned from scratch, can render novel views with as few as one observed image when pre-trained on a multi-view dataset, and produces plausible completions of completely unobserved regions. Our project website is available at https://www.ajayj.com/dietnerf.

1. Introduction

In the novel view synthesis problem, we seek to rerender a scene from arbitrary viewpoint given a set of sparsely sampled viewpoints. View synthesis is a challenging problem that requires some degree of 3D reconstruction in addition to high-frequency texture synthesis. Recently, great progress has been made on high-quality view synthesis when many observations are available. A popular approach is to use Neural Radiance Fields (NeRF) [25] to estimate a continuous neural scene representation from image observations. During training on a particular scene, the representation is rendered from observed viewpoints using volumetric ray casting to compute a reconstruction loss. At test time, NeRF can be rendered from novel viewpoints by the same procedure. While conceptually very simple, NeRF can learn high-frequency view-dependent scene appearances and accurate geometries that allow for high-quality rendering.

Still, NeRF is estimated per-scene, and cannot benefit from prior knowledge acquired from other images and objects. Because of the lack of prior knowledge, NeRF requires

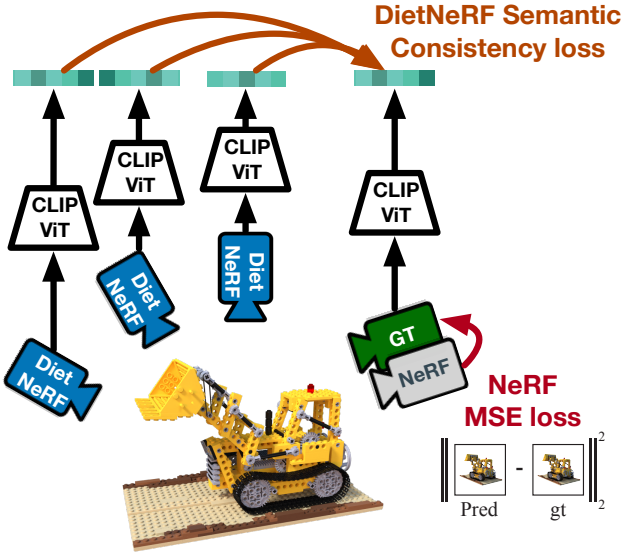

Figure 1. Neural Radiance Fields are trained to represent a scene by supervising renderings from the *same pose* as ground-truth observations (**MSE loss**). However, when only a few views are available, the problem is underconstrained. NeRF often finds degenerate solutions unless heavily regularized. Based on the principle that **"a bulldozer is a bulldozer from any perspective"**, our proposed DietNeRF supervises the radiance field from arbitrary poses (**DietNeRF cameras**). This is possible because we compute a **semantic consistency loss** in a feature space capturing high-level scene attributes, not in pixel space. We extract semantic representations of renderings using the CLIP Vision Transformer [28], then maximize similarity with representations of ground-truth views. In effect, we use prior knowledge about scene semantics learned by *single-view* 2D image encoders to constrain a 3D representation.

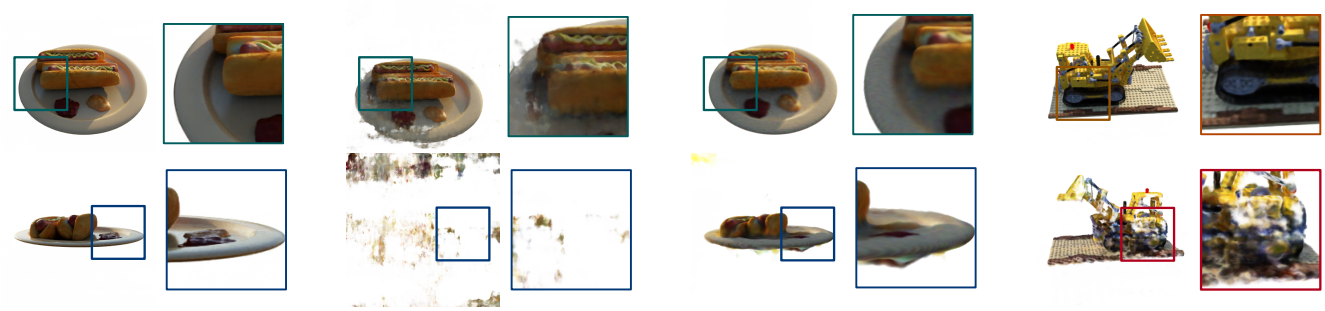

Figure 2. **Few-shot view synthesis is a challenging problem for Neural Radiance Fields.** **(A)** When we have 100 observations of an object from uniformly sampled poses, NeRF estimates a detailed and accurate representation that allows for high-quality view synthesis purely from multi-view consistency. **(B)** However, with only 8 views, the same NeRF overfits by placing the object in the near-field of the training cameras, leading to misplaced objects at poses near training cameras and degeneracies at novel poses. **(C)** We find that NeRF can converge when regularized, simplified, tuned and manually reinitialized, but no longer captures fine details. **(D)** Finally, without prior knowledge about similar objects, single-scene view synthesis cannot plausibly complete unobserved regions, such as the left side of an object seen from the right. In this work, we find that **these failures occur because NeRF is only supervised from the sparse training poses.**

a large number of input views to reconstruct a given scene at high-quality. Given 8 views, Figure 2B shows that novel views rendered with the full NeRF model contain many artifacts because the optimization finds a degenerate solution that is only accurate at observed poses. We find that the core issue is that prior 3D reconstruction systems based on rendering losses are *only supervised at known poses*, so they overfit when few poses are observed. Regularizing NeRF by simplifying the architecture avoids the worst artifacts, but comes at the cost of fine-grained detail.

Further, prior knowledge is needed when the scene reconstruction problem is underdetermined. 3D reconstruction systems struggle when regions of an object are never observed. This is particularly problematic when rendering an object at significantly different poses. When rendering a scene with an extreme baseline change, unobserved regions during training become visible. A view synthesis system should generate plausible missing details to fill in the gaps. Even a regularized NeRF learns poor extrapolations to unseen regions due to its lack of prior knowledge (Figure 2D).

Recent work trained NeRF on *multi-view* datasets of similar scenes [44, 38, 32, 37, 42] to bias reconstructions of novel scenes. Unfortunately, these models often produce blurry images due to uncertainty, or are restricted to a single object category such as ShapeNet classes as it is challenging to capture large, diverse, multi-view data.

In this work, we exploit the consistency principle that *"a bulldozer is a bulldozer from any perspective"*: objects share high-level semantic properties between their views. Image recognition models learn to extract many such high-level semantic features including object identity. We transfer prior knowledge from pre-trained image encoders learned on highly diverse 2D *single-view* image data to the view synthesis problem. In the single-view setting, such encoders are frequently trained on millions of realistic images like ImageNet [7]. CLIP is a recent multi-modal encoder that is trained to match images with captions in a massive web scrape containing 400M images [28]. Due to the diversity of its data, CLIP showed promising zero- and few-shot transfer performance to image recognition tasks. We find that CLIP and ImageNet models also contain prior knowledge useful for novel view synthesis.

We propose DietNeRF, a neural scene representation based on NeRF that can be estimated from only a few photos, and can generate views with unobserved regions. In addition to minimizing NeRF's mean squared error losses at known poses in pixel-space, DietNeRF penalizes a *semantic consistency* loss. This loss matches the final activations of CLIP's Vision Transformer [9] between ground-truth images and rendered images at *different* poses, allowing us to supervise the radiance field from arbitrary poses. In experiments, we show that DietNeRF learns realistic reconstructions of objects with as few as 8 views without simplifying the underlying volumetric representation, and can even produce reasonable reconstructions of completely occluded regions. To generate novel views with as few as 1 observation, we fine-tune pixelNeRF [44], a generalizable scene representation, and improve perceptual quality.

2. Background on Neural Radiance Fields

A plenoptic function, or light field, is a five-dimensional function that describes the light radiating from every point in every direction in a volume such as a bounded scene. While explicitly storing or estimating the plenoptic function at high resolution is impractical due to the dimensionality of the

input, Neural Radiance Fields [25] parameterize the function with a continuous neural network such as a multi-layer perceptron (MLP). A Neural Radiance Field (NeRF) model is a five-dimensional function $f_\theta(\mathbf{x}, \mathbf{d}) = (\mathbf{c}, \sigma)$ of spatial position $\mathbf{x} = (x, y, z)$ and viewing direction (θ, ϕ), expressed as a 3D unit vector $\mathbf{d}$. NeRF predicts the RGB color $\mathbf{c}$ and differential volume density σ from these inputs. To encourage view-consistency, the volume density only depends on $\mathbf{x}$, while the color also depends on viewing direction $\mathbf{d}$ to capture viewpoint dependent effects like specular reflections. Images are rendered from a virtual camera at any position by integrating color along rays cast from the observer according to volume rendering [18]:

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

where the ray originating at the camera origin $\mathbf{o}$ follows path $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, and the transmittance $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$ weights the radiance by the probability that the ray travels from the image plane at t_n to t unobstructed. To approximate the integral, NeRF employs a hierarchical sampling algorithm to select function evaluation points near object surfaces along each ray. NeRF separately estimates two MLPs, a coarse network and a fine network, and uses the coarse network to guide sampling along the ray for more accurately estimating (1). The networks are trained from scratch on *each scene* given tens to hundreds of photos from various perspectives. Given observed multi-view training images $\{I_i\}$ of a scene, NeRF uses COLMAP SfM [31] to estimate camera extrinsics (rotations and origins) $\{\mathbf{p}_i\}$, creating a posed dataset $\mathcal{D} = \{(I_i, \mathbf{p}_i)\}$.

3. NeRF Struggles at Few-Shot View Synthesis

View synthesis is a challenging problem when a scene is only sparsely observed. Systems like NeRF that train on individual scenes especially struggle without prior knowledge acquired from similar scenes. We find that NeRF fails at few-shot novel view synthesis in several settings.

NeRF overfits to training views Conceptually, NeRF is trained by mimicking the image-formation process at observed poses. The radiance field can be estimated repeatedly sampling a training image and pose $(I, \mathbf{p}_i)$, rendering an image $\hat{I}_{\mathbf{p}_i}$ from the **same pose** by volume integration (1), then minimizing the mean-squared error (MSE) between the images, which should align pixel-wise:

$$\mathcal{L}_{\text{full}}(I, \hat{I}_{\mathbf{p}_i}) = \frac{1}{HW}\|I - \hat{I}_{\mathbf{p}_i}\|_2^2 \tag{2}$$

In practice, NeRF samples a smaller batch of rays across all training images to avoid the computational expense of rendering full images during training. Given subsampled rays $\mathcal{R}$ cast from the training cameras, NeRF minimizes:

$$\mathcal{L}_{\text{MSE}}(\mathcal{R}) = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}} \|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|_2^2 \tag{3}$$

With many training views, $\mathcal{L}_{\text{MSE}}$ provides training signal to f_θ densely in the volume and does not overfit to individual training views. Instead, the MLP recovers accurate textures and occupancy that allow interpolations to new views (Figure 2A). Radiance fields with sinusoidal positional embeddings are quite effective at learning high-frequency functions [37], which helps the MLP represent fine details.

Unfortunately, this high-frequency representational capacity allows NeRF to overfit to each input view when only a few are available. $\mathcal{L}_{\text{MSE}}$ can be minimized by packing the reconstruction $\hat{I}_{\mathbf{p}}$ of training view $(I, \mathbf{p})$ close to the camera. Fundamentally, the plenoptic function representation suffers from a near-field ambiguity [45] where distant cameras each observe significant regions of space that no other camera observes. In this case, the optimal scene representation is underdetermined. Degenerate solutions can also exploit the view-dependence of the radiance field. Figure 2B shows novel views from the same NeRF trained on 8 views. While a rendered view from a pose near a training image has reasonable textures, it is skewed incorrectly and has cloudy artifacts from incorrect geometry. As the geometry is not estimated correctly, a distant view contains almost none of the correct information. High-opacity regions block the camera. Without supervision from any nearby camera, opacity is sensitive to random initialization.

Regularization fixes geometry, but hurts fine-detail High-frequency artifacts such as spurious opacity and rapidly varying colors can be avoided in some cases by regularizing NeRF. We simplify the NeRF architecture by removing hierarchical sampling and learning only a single MLP, and reducing the maximum frequency positional embedding in the input layer. This biases NeRF toward lower frequency solutions, such as placing content in the center of the scene farther from the training cameras. We also can address some few-shot optimization challenges by lowering the learning rate to improve initial convergence, and manually restarting training if renderings are degenerate. Figure 2C shows that these regularizers successfully allow NeRF to recover plausible object geometry. However, high-frequency, fine details are lost compared to 2A.

No prior knowledge, no generalization to unseen views As NeRF is estimated from scratch per-scene, it has no prior knowledge about natural objects such as common symmetries and object parts. In Figure 2D, we show that NeRF trained with 14 views of the right half of a Lego vehicle generalizes poorly to its left side. We regularized NeRF to remove high-opacity regions that originally blocked the left side entirely. Even so, the essential challenge is that NeRF receives no supervisory signal from $\mathcal{L}_{\text{MSE}}$ to the unob-

served regions, and instead relies on the inductive bias of the MLP for any inpainting. We would like to introduce prior knowledge that allows NeRF to exploit bilateral symmetry for plausible completions.

4. Semantically Consistent Radiance Fields

Motivated by these challenges, we introduce the DietNeRF scene representation. DietNeRF uses prior knowledge from a pre-trained image encoder to guide the NeRF optimization process in the few-shot setting.

4.1. Semantic consistency loss

DietNeRF supervises f_θ at arbitrary camera poses during training with a semantic loss. While pixel-wise comparison between ground-truth observed images and rendered images with $\mathcal{L}_{\text{MSE}}$ is only useful when the rendered image is aligned with the observed pose, humans are easily able to detect whether two images are views of the same object from semantic cues. We can in general compare a *representation* of images captured from different viewpoints:

$$\mathcal{L}_{\text{SC},\ell_2}(I, \hat{I}) = \frac{\lambda}{2}\|\phi(I) - \phi(\hat{I})\|_2^2 \tag{4}$$

If $\phi(x) = x$, Eq. (4) reduces to $\mathcal{L}_{\text{full}}$ up to a scaling factor. However, the identity mapping is view-dependent. We need a representation that is similar across views of the same object and captures important high-level semantic properties like object class. We evaluate the utility of two sources of supervision for representation learning. First, we experiment with the recent CLIP model pre-trained for multi-modal language and vision reasoning with contrastive learning [28]. We then evaluate visual classifiers pre-trained on labeled ImageNet images [9]. In both cases, we use similar Vision Transformer (ViT) architectures.

A Vision Transformer is appealing because its performance scales very well to large amounts of 2D data. Training on a large variety of images allows the network to encounter multiple views of an object class over the course of training without explicit multi-view data capture. It also allows us to transfer the visual encoder to diverse objects of interest in graphics applications, unlike prior class-specific reconstruction work that relies on homogeneous datasets [3, 19]. ViT extracts features from non-overlapping image patches in its first layer, then aggregates increasingly abstract representations with Transformer blocks based on global self-attention [41] to produce a single, global embedding vector. ViT outperformed CNN encoders in our early experiments.

In practice, CLIP produces normalized image embeddings. When $\phi(\cdot)$ is a unit vector, Eq. (4) simplifies to cosine similarity up to a constant and a scaling factor that can be absorbed into the loss weight λ:

$$\mathcal{L}_{\text{SC}}(I, \hat{I}) = \lambda\phi(I)^T\phi(\hat{I}) \tag{5}$$

Algorithm 1: Training DietNeRF on a single scene

Data: Observed views $\mathcal{D} = \{(I, \mathbf{p})\}$, semantic embedding function $\phi(\cdot)$, pose distribution π, consistency interval K, weight λ, rendering size, batch size $|\mathcal{R}|$, lr η_{it}
Result: Trained Neural Radiance Field $f_\theta(\cdot, \cdot)$
Initialize NeRF $f_\theta(\cdot, \cdot)$;
Pre-compute target embeddings $\{\phi(I) : I \in \mathcal{D}\}$;
for *it from 1 to num_iters* **do**
  Sample ray batch $\mathcal{R}$, ground-truth colors $\mathbf{C}(\cdot)$;
  Render rays $\hat{\mathbf{C}}(\cdot)$ by (1);
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{MSE}}(\mathcal{R}, \mathbf{C}, \hat{\mathbf{C}})$;
  if *it* $\%$ $K = 0$ **then**
    Sample target image, pose $(I, \mathbf{p}) \sim \mathcal{D}$;
    Sample source pose $\hat{\mathbf{p}} \sim \pi$;
    Render image $\hat{I}$ from pose $\hat{\mathbf{p}}$;
    $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{\text{SC}}(I, \hat{I})$;
  end
  Update parameters: $\theta \leftarrow Adam(\theta, \eta_{it}, \nabla_\theta\mathcal{L})$;
end

We refer to $\mathcal{L}_{\text{SC}}$ (5) as a *semantic consistency* loss because it measures the similarity of high-level semantic features between observed and rendered views. In principle, semantic consistency is a very general loss that can be applied to any 3D reconstruction system based on differentiable rendering.

4.2. Interpreting representations across views

The pre-trained CLIP model that we use is trained on hundreds of millions of images with captions of varying detail. Image captions provide rich supervision for image representations. On one hand, short captions express semantically sparse learning signal as a flexible way to express labels [8]. For example, the caption "A photo of hotdogs" describes Fig. 2A. Language also provides semantically dense learning signal by describing object properties, relationships and appearances [8] such as the caption "Two hotdogs on a plate with ketchup and mustard". To be predictive of such captions, an image representation must capture some high-level semantics that are stable across viewpoints. Concurrently, [10] found that CLIP representations capture visual attributes of images like art style and colors, as well as high-level semantic attributes including object tags and categories, facial expressions, typography, geography and brands.

In Figure 3, we measure the pairwise cosine similarity between CLIP representations of views circling an object. We find that pairs of views have highly similar CLIP representations, even for diametrically opposing cameras. This suggests that large, diverse single-view datasets can induce useful representations for multi-view applications.

4.3. Pose sampling distribution

We augment the NeRF training loop with $\mathcal{L}_{\text{SC}}$ minimization. Each iteration, we compute $\mathcal{L}_{\text{SC}}$ between a random training image sampled from the observation dataset

$I \sim \mathcal{D}$ and rendered image $\hat{I}_{\mathbf{p}}$ from random pose $\mathbf{p} \sim \pi$. For bounded scenes like NeRF's Realistic Synthetic scenes where we are interested in 360° view synthesis, we define the pose sampling distribution π to be a uniform distribution over the upper hemisphere, with radius sampled uniformly in a bounded range. For unbounded forward-facing scenes or scenes where a pose sampling distribution is difficult to define, we interpolate between three randomly sampled known poses $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3 \sim \mathcal{D}$ with pairwise interpolation weights $\alpha_1, \alpha_2 \sim \mathcal{U}(0, 1)$.

4.4. Improving efficiency and quality

Volume rendering is computationally intensive. Computing a pixel's color evaluates NeRF's MLP f_θ at many points along a ray. To improve the efficiency of DietNeRF during training, we render images for semantic consistency at low resolution, requiring only 15-20% of the rays as a full resolution training image. Rays are sampled on a strided grid across the full extent of the image plane, ensuring that objects are mostly visible in each rendering. We found that sampling poses from a continuous distribution was helpful to avoid aliasing artifacts when training at a low resolution.

In experiments, we found that $\mathcal{L}_{\text{SC}}$ converges faster than $\mathcal{L}_{\text{MSE}}$ for many scenes. We hypothesize that the semantic consistency loss encourages DietNeRF to recover plausible scene geometry early in training, but is less helpful for reconstructing fine-grained details due to the relatively low dimensionality of the ViT representation $\phi(\cdot)$. We exploit the rapid convergence of $\mathcal{L}_{\text{SC}}$ by only minimizing $\mathcal{L}_{\text{SC}}$ every k iterations. DietNeRF is robust to the choice of k, but a value between 10 and 16 worked well in our experiments. StyleGAN2 [20] used a similar strategy for efficiency, referring to periodic application of a loss as *lazy regularization*.

As backpropagation through rendering is memory intensive with reverse-mode automatic differentiation, we render images for $\mathcal{L}_{\text{SC}}$ with mixed precision computation and evaluate $\phi(\cdot)$ at half-precision. We delete intermediate MLP activations during rendering and rematerialize them during the backward pass [6, 15]. All experiments use a single 16 GB NVIDIA V100 or 11 GB 2080 Ti GPU.

Since $\mathcal{L}_{\text{SC}}$ converges before $\mathcal{L}_{\text{MSE}}$, we found it helpful to fine-tune DietNeRF with $\mathcal{L}_{\text{MSE}}$ alone for 20-70k iterations to refine details. Alg. 1 details our overall training process.

5. Experiments

In experiments, we evaluate the quality of novel views synthesized by DietNeRF and baselines for both synthetically rendered objects and real photos of multi-object scenes. (1) We evaluate training *from scratch* on a specific scene with 8 views §5.1. (2) We show that DietNeRF improves perceptual quality of view synthesis from *only a single real photo* §5.2. (3) We find that DietNeRF can reconstruct regions that are never observed §5.3, and finally (4) run ablations §5.4.

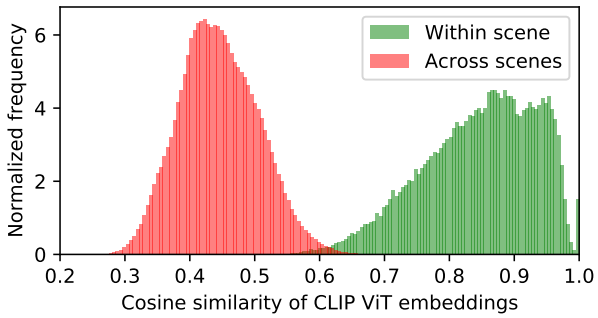

Figure 3. CLIP's Vision Transformer learns low-dimensional image representations through language supervision. We find that these representations transfer well to multi-view 3D settings. We sample pairs of ground-truth views of the same scene and of different scenes from NeRF's Realistic Synthetic object dataset, then compute a histogram of representation cosine similarity. Even though camera poses vary dramatically (views are sampled from the upper hemisphere), views within a scene have similar representations (**green**). Across scenes, representations have low similarity (**red**)

Datasets The Realistic Synthetic benchmark of [24] includes detailed multi-view renderings of 8 realistic objects with view-dependent light transport effects. We also benchmark on the DTU multi-view stereo (MVS) dataset [16] used by pixelNeRF [44]. DTU is a challenging dataset that includes sparsely sampled real photos of physical objects.

Low-level full reference metrics Past work evaluates novel view quality with respect to ground-truth from the same pose with Peak Signal-to-Noise Ratio (**PSNR**) and Structural Similarity Index Measure (**SSIM**) [35]. PSNR expresses mean-squared error in log space. However, SSIM often disagrees with human judgements of similarity [46].

Perceptual metrics Deep CNN activations mirror aspects of human perception. NeRF measures perceptual image quality using **LPIPS** [46], which computes MSE between normalized features from all layers of a pre-trained VGG encoder [33]. Generative models also measure sample quality with feature space distances. The Fréchet Inception Distance (**FID**) [12] computes the Fréchet distance between Gaussian estimates of penultimate Inception v3 [36] features for real and fake images. However, FID is a biased metric at low sample sizes. We adopt the conceptually similar Kernel Inception Distance (**KID**), which measures the MMD between Inception features and has an unbiased estimator [2, 26]. All metrics use a different architecture and data than our CLIP ViT encoder.

5.1. Realistic Synthetic scenes from scratch

NeRF's Realistic Synthetic dataset includes 8 detailed synthetic objects with 100 renderings from virtual cameras arranged randomly on a hemisphere pointed inward. To test few-shot performance, we *randomly* sample a training sub-

Table 1. Quality metrics for novel view synthesis on subsampled splits of the Realistic Synthetic dataset [25]. We randomly sample 8 views from the available 100 ground truth training views to evaluate how DietNeRF performs with limited observations.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
NeRF	14.934	0.687	0.318	228.1	0.076
NV	17.859	0.741	0.245	239.5	0.117
Simplified NeRF	20.092	0.822	0.179	189.2	0.047
DietNeRF (ours)	23.147	0.866	0.109	74.9	0.005
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft	**23.591**	**0.874**	**0.097**	**72.0**	**0.004**
NeRF, 100 views	**31.153**	**0.954**	**0.046**	**50.5**	**0.001**

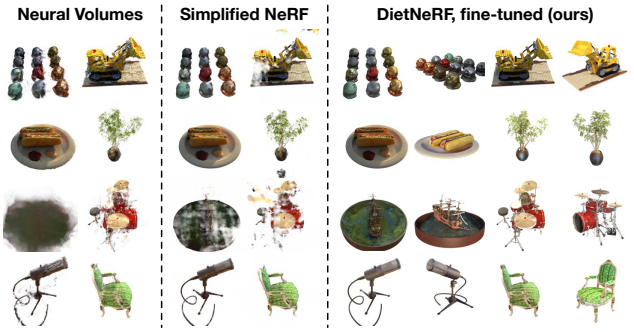

Figure 4. Novel views synthesized from eight observations of scenes in the Realistic Synthetic dataset.

set of 8 images from each scene. Tab. 1 shows results. The original NeRF achieves much poorer quality with 8 images than with the full 100 image dataset. Neural Volumes [23] performs better as it tightly constrains the size of the scene's bounding box and explicitly regularizes its scene representation with a penalty on spatial gradients of voxel opacity and a Beta prior on image opacity. This avoids the worst artifacts, but reconstructions are still low-quality. Simplifying NeRF and tuning for each individual scene also regularizes the representation and helps convergence (+5.1 PSNR over the full NeRF). The best performance is achieved by regularizing with DietNeRF's $\mathcal{L}_{\text{SC}}$ loss. Additionally, fine-tuning with $\mathcal{L}_{\text{MSE}}$ even further improves quality, for a total improvement of +8.5 PSNR, -0.2 LPIPS, and -156 FID over NeRF. This shows that semantic consistency is a valuable prior for high-quality few-shot view synthesis. Fig. 4 visualizes results.

5.2. Single-view synthesis by fine-tuning

NeRF only uses observations during training, not inference, and uses no auxiliary data. Accurate 3D reconstruction from a single view is not possible purely from $\mathcal{L}_{\text{MSE}}$, so NeRF performs poorly in the single-view setting (Table 2).

To perform single- or few-shot view synthesis, pixelNeRF [44] learns a ResNet-34 encoder and a feature-conditioned neural radiance field on a multi-view dataset of similar scenes. The encoder learns priors that generalize

Table 2. **Single-view novel view synthesis on the DTU dataset**. NeRF and pixelNeRF PSNR, SSIM and LPIPS results from [44]. Finetuning pixelNeRF with our semantic consistency loss (DietPixelNeRF) improves perceptual quality by the deep perceptual LPIPS, FID and KID evaluation metrics, but can degrade PSNR and SSIM which are local pixel-aligned metrics due to geometric defects.

Method	PSNR	SSIM	LPIPS	FID	KID
NeRF	8.000	0.286	0.703	—	—
pixelNeRF	15.550	0.537	0.535	266.1	0.166
pixelNeRF, $\mathcal{L}_{\text{MSE}}$ ft	**16.048**	**0.564**	0.515	265.2	0.159
DietPixelNeRF	14.242	0.481	**0.487**	**190.7**	**0.066**

Table 3. **Extrapolation metrics.** Novel view synthesis with observations of **only one side** of the Realistic Synthetic Lego scene.

Views	Method	PSNR ↑	SSIM ↑	LPIPS ↓
14	NeRF	19.662	0.799	0.202
14	Simplified NeRF	21.553	0.818	0.160
14	DietNeRF (ours)	20.753	0.810	0.157
14	DietNeRF + $\mathcal{L}_{\text{MSE}}$ ft	**22.211**	**0.824**	**0.143**
100	NeRF [25]	**31.618**	**0.965**	**0.033**

to new single-view scenes. Table 2 shows that pixelNeRF significantly outperforms NeRF given a single photo of a held-out scene. However, novel views are blurry and unrealistic (Figure 5). We propose to fine-tune pixelNeRF on a single scene using $\mathcal{L}_{\text{MSE}}$ alone or using both $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{SC}}$. Fine-tuning per-scene with MSE improves local image quality metrics, but only slightly helps perceptual metrics. Figure 6 shows that pixel-space MSE fine-tuning from one view mostly only improves quality for that view.

We refer to fine-tuning with both losses for a short period as DietPixelNeRF. Qualitatively, DietPixelNeRF has significantly sharper novel views (Fig. 5, 6). DietPixelNeRF outperforms baselines on perceptual LPIPS, FID, and KID metrics (Tab. 2). For the very challenging single-view setting, ground-truth novel views will contain content that is completely occluded in the input. Because of uncertainty, blurry renderings will outperform sharp but incorrect renderings on average error metrics like MSE and PSNR. Arguably, perceptual quality and sharpness are better metrics than pixel error for graphics applications like photo editing and virtual reality as plausibility is emphasized.

5.3. Reconstructing unobserved regions

We evaluate whether DietNeRF produces plausible completions when the reconstruction problem is underdetermined. For training, we sample 14 nearby views of the right side of the Realistic Synthetic Lego scene (Fig. 7, right). Narrow baseline multi-view capture rigs are less costly than 360° captures, and support unbounded scenes. However, narrow-baseline observations suffer from occlusions: the

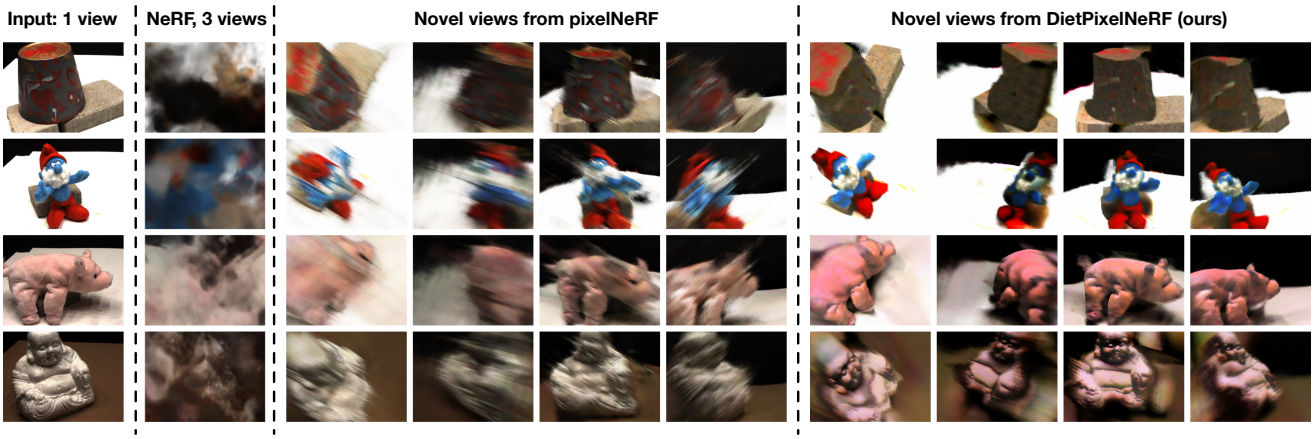

Figure 5. **Novel views synthesized from a *single input image*** from the DTU object dataset. Even with 3 input views, NeRF [25] fails to learn accurate geometry or textures (reprinted from [44]). While pixelNeRF [44] has mostly consistent object geometry as the camera pose is varied, renderings are blurry and contain artifacts like inaccurate placement of density along the observed camera's z-axis. In contrast, fine-tuning with DietNeRF (DietPixelNeRF) learns realistic textures visually consistent with the input image, though some geometric defects are present due to the ambiguous nature of the view synthesis problem.

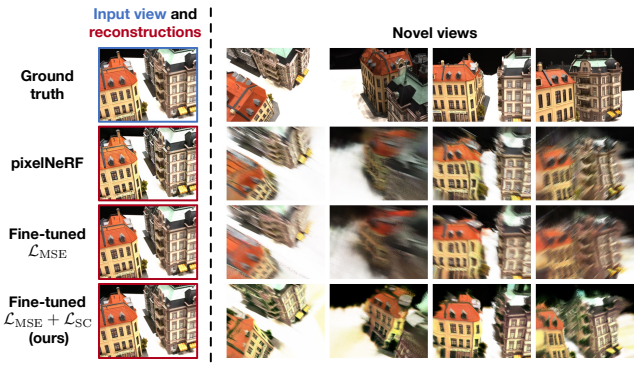

Figure 6. **Semantic consistency improves perceptual quality.** Fine-tuning pixelNeRF with $\mathcal{L}_{\text{MSE}}$ slightly improves a rendering of the input view, but does not remove most perceptual flaws like blurriness in *novel views*. Fine-tuning with both $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{SC}}$ (DietPixelNeRF, bottom) improves sharpness of all views.

left side of the Lego bulldozer is unobserved. NeRF fails to reconstruct this side of the scene, while our Simplified NeRF learns unrealistic deformations and incorrect colors (Fig. 7, left). Remarkably, DietNeRF learns quantitatively (Tab. 3) and qualitatively more accurate colors in the missing regions, suggesting the value of semantic image priors for sparse reconstruction problems. We exclude FID and KID since a single scene has too few samples for an accurate estimate.

5.4. Ablations

Choosing an image encoder Table 4 shows quality metrics with different semantic encoder architectures and pre-training datasets. We evaluate on the Lego scene with 8 views. Large ViT models (ViT L) do not improve results over the base ViT B. CLIP's ViT B/32 offers a +1.8 PSNR improvement over an ImageNet model, suggesting that data diversity and language supervision is helpful for 3D tasks. Still, both induce useful representations that transfer to view synthesis. Using CLIP's ResNet-50 gives PSNR 17.15. DietNeRF performs best with ViTs, but a CNN can help too.

Table 4. **Ablating supervision and architectural parameters for the ViT image encoder $\phi(\cdot)$ used to compare image features.** Metrics are measured on the Realistic Synthetic Lego scene.

Semantic image encoder	PSNR ↑	SSIM ↑	LPIPS ↓
ImageNet ViT L/16, 384^2	21.501	0.809	0.167
ImageNet ViT L/32, 384^2	20.498	0.801	0.174
ImageNet ViT B/32, 224^2	22.059	0.836	0.131
CLIP ViT B/32, 224^2	**23.896**	**0.863**	**0.110**

Table 5. **Varying the number of iterations that DietNeRF is fine-tuned with $\mathcal{L}_{\text{MSE}}$** on Realistic Synthetic scenes. All models are initially trained for 200k iterations with $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{SC}}$. Further minimizing $\mathcal{L}_{\text{MSE}}$ is helpful, but the model can overfit.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
DietNeRF, no fine-tuning	23.147	0.866	0.109
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft 10k iters	23.524	0.872	0.101
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft 50k iters	**23.591**	**0.874**	**0.097**
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft 100k iters	23.521	**0.874**	**0.097**
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft 200k iters	23.443	0.872	0.098

Varying $\mathcal{L}_{\text{MSE}}$ fine-tuning duration Fine-tuning DietNeRF with $\mathcal{L}_{\text{MSE}}$ can improve quality by better reconstructing fine-details. In Table 5, we vary the number of iterations of fine-tuning for the Realistic Synthetic scenes with 8 views. Fine-tuning for up to 50k iterations is helpful, but reduces performance with longer optimization. It is possible that the model starts overfitting to the 8 input views.

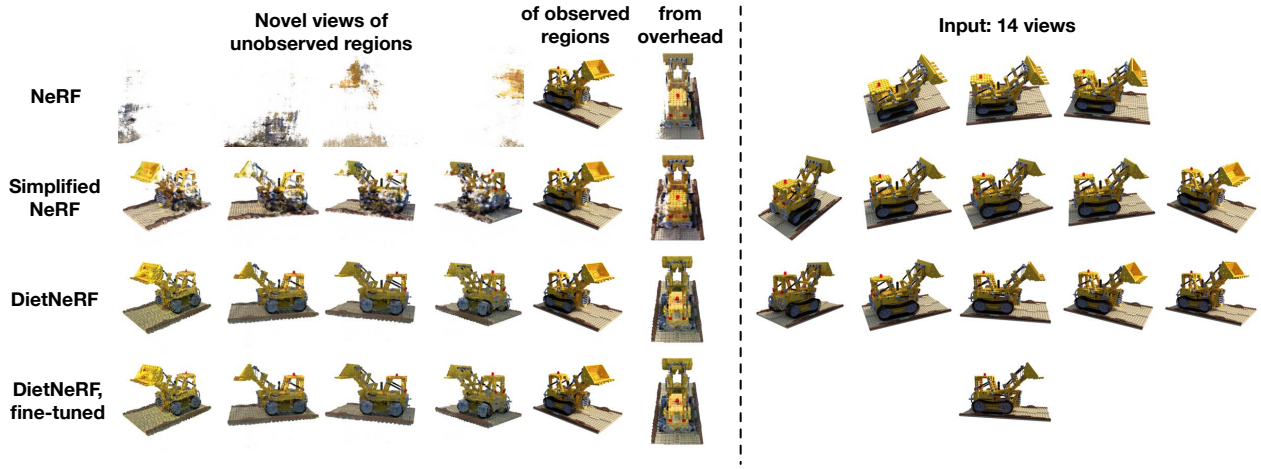

Figure 7. **Renderings of occluded regions during training.** 14 images of the right half of the Realistic Synthetic lego scene are used to estimate radiance fields. NeRF either learns high-opacity occlusions blocking the left of the object, or fails to generalize properly to the unseen left side. In contrast, DietNeRF fills in details for a reconstruction that is mostly consistent with the observed half.

6. Related work

Few-shot radiance fields Several works condition NeRF on latent codes describing scene geometry or appearance rather than estimating NeRF per scene [32, 38, 44]. An image encoder and radiance field decoder are learned on a multi-view dataset of similar objects or scenes ahead of time. At test time, on a new scene, novel viewpoints are rendered using the decoder conditioned on encodings of a few observed images. GRAF renders patches of the scene every iteration to supervise the network with a discriminator [32]. Concurrently, IBRNet [42] also fine-tunes a latent-conditioned radiance field on a specific scene using NeRF's reconstruction loss, but needed at least 50 views. Rather than generalizing through a shared encoder and decoder, [37] meta-learns radiance field weights that can be adapted to specific scenes in a few gradient steps. Meta-learning improves few-view performance. Similarly, [34] meta-learns a signed distance field for shape representation problems. Much literature studies single-view reconstruction with explicit 3D representations. Notable recent examples include voxel [39], mesh [13] and point-cloud [43] approaches.

Novel view synthesis, image-based rendering Neural Volumes [23] proposes a VAE [21, 29] encoder-decoder architecture to predict a volumetric representation of a scene from posed image observations. NV uses priors as auxiliary objectives like DietNeRF, but penalizes opacity based on geometric intuitions rather than RGB image semantics. TBNs [27] learn an autoencoder with a 3D latent that can be rotated to render new perspectives for a single-category. SRNs [35] fit a continuous representation to a scene and generalize to novel single-category objects if trained on a large multi-view dataset. It can be extended to predict per-point semantic segmentation maps [22]. Local Light Field Fusion [24] estimates and blends multiple MPI representations for each scene. Free View Synthesis [30] uses geometric approaches to improve unbounded in-the-wild scenes. NeRF++ [45] also improves unbounded scenes using multiple NeRF models and changing NeRF's parameterization.

Semantic representation learning Representation learning with deep supervised and unsupervised approaches has a long history [1]. Without labels, generative models can learn useful representations for recognition [4], but self-supervised models like CPC [40, 11] tend to be more parameter efficient. Contrastive methods including CLIP learn visual representations by matching similar pairs of items, such as captions and images [28, 17], augmentated variants of an image [5], or video patches across frames [14].

7. Conclusions

Our results suggest that single-view 2D representations transfer effectively to underconstrained 3D reconstruction problems such as volumetric novel view synthesis. While pre-trained image representations have certainly been transferred to 3D vision applications in the past by fine-tuning, the recent emergence of visual models trained on enormous 100M+ image datasets like CLIP have enabled surprisingly effective few-shot transfer. We exploited this transferable prior knowledge to solve optimization issues and to cope with partial observability in the NeRF family of scene representations, with notable improvements in perceptual quality. In the future, we believe "diet-friendly" few-shot transfer will play a greater role in a wide range of 3D applications.

Acknowledgements Our work is supported by the NSF GRFP (grant DGE-1752814) and Berkeley Deep Drive. We thank Alexei Efros, Paras Jain, Aditi Jain, Angjoo Kanazawa, Aravind Srinivas and Alex Yu for helpful feedback.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 8

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 5

[3] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):232–244, 2013. 4

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020. 8

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 8

[6] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009. 2

[8] Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 4

[10] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 2021. https://distill.pub/2021/multimodal-neurons. 4

[11] Olivier J Henaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 13–18 Jul 2020. 8

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. 5

[13] Ronghang Hu and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. *arXiv preprint arXiv:2012.09854*, 2020. 8

[14] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. *Advances in Neural Information Processing Systems*, 2020. 8

[15] Paras Jain, Ajay Jain, Aniruddha Nrusimha, Amir Gholami, Pieter Abbeel, Joseph Gonzalez, Kurt Keutzer, and Ion Stoica. Checkmate: Breaking the memory wall with optimal tensor rematerialization. In *Proceedings of Machine Learning and Systems*, volume 2, pages 497–511, 2020. 5

[16] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 5

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 8

[18] James T. Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '84, page 165–174, New York, NY, USA, 1984. Association for Computing Machinery. 3

[19] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 4

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 8

[22] Amit Kohli, Vincent Sitzmann, and Gordon Wetzstein. Semantic implicit neural scene representations with semi-supervised training. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 423–433, Los Alamitos, CA, USA, nov 2020. IEEE Computer Society. 8

[23] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 6, 8

[24] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 5, 8

[25] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 6, 7

[26] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch, 2020. Version: 0.2.0, DOI: 10.5281/zenodo.3786540. 5

[27] Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. *The IEEE International Conference on Computer Vision (ICCV)*, Nov 2019. 8

[28] 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 *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. 1, 2, 4, 8

[29] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. 8

[30] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 623–640, Cham, 2020. Springer International Publishing. 8

[31] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[32] 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, 8

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[34] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions, 2020. 8

[35] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 5, 8

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,*, 2016. 5

[37] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 8

[38] Alex Trevithick and Bo Yang. GRF: Learning a general radiance field for 3d scene representation and rendering. In *arXiv:2010.04595*, 2020. 2, 8

[39] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 8

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 4

[42] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 8

[43] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8

[44] 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 (CVPR)*, 2021. 2, 5, 6, 7, 8

[45] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 3, 8

[46] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5