SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image

Dejia Xu[1*], Yifan Jiang[1*], Peihao Wang[1], Zhiwen Fan[1]
Humphrey Shi[2,3,4], Zhangyang Wang[1]

[1]The University of Texas at Austin, [2]UIUC,
[3]University of Oregon, [4]Picsart AI Research

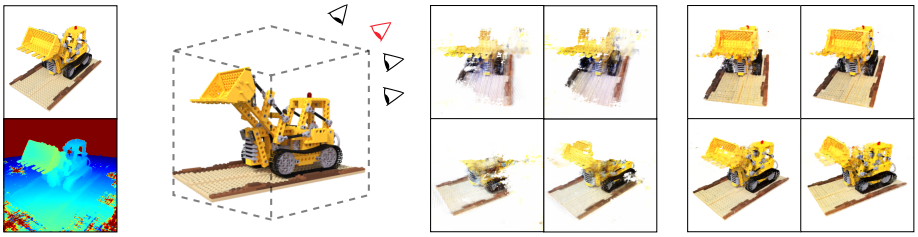

Fig. 1: Given only a single reference view as input, our novel semi-supervised framework effectively trains a neural radiance field. In contrast, previous method [9] shows inconsistent geometry when synthesizing novel views.

Abstract. Despite the rapid development of Neural Radiance Field (NeRF), the necessity of dense covers largely prohibits its wider applications. While several recent works have attempted to address this issue, they either operate with sparse views (yet still, a few of them) or on simple objects/scenes. In this work, we consider a more ambitious task: training neural radiance field, over realistically complex visual scenes, by "looking only once", i.e., using only a **single** view. To attain this goal, we present a *Single View NeRF* (**SinNeRF**) framework consisting of thoughtfully designed semantic and geometry regularizations. Specifically, SinNeRF constructs a semi-supervised learning process, where we introduce and propagate geometry pseudo labels and semantic pseudo labels to guide the progressive training process. Extensive experiments are conducted on complex scene benchmarks, including NeRF synthetic dataset, Local Light Field Fusion dataset, and DTU dataset. We show that even without pre-training on multi-view datasets, SinNeRF can yield photo-realistic novel-view synthesis results. Under the single image setting, SinNeRF significantly outperforms the current state-of-the-art NeRF baselines in all cases. Project page: `https://vita-group.github.io/SinNeRF/`

⋆ Equal contribution

1 Introduction

Synthesizing photo-realistic images has been one of the most essential goals in the area of computer vision. Recently, the field of novel view synthesis has gained tremendous popularity with the success of coordinate-based neural networks. Neural radiance field (NeRF) [31], as an effective scene representation, has prevailed among image-based rendering approaches.

Despite its great success, NeRF is impeded by the stringent requirement of the dense views captured from different angles and the corresponding camera poses. As has been implied by recent literature [34]., training a neural radiance field without sufficient views will end up with drastic performance degradation, including incorrect geometry and blurry appearance. Meanwhile, it could be challenging or even infeasible in real-world scenarios to collect a sufficiently dense coverage of views for specific applications such as AR/VR or autonomous driving. Motivated by this, many researchers attempt to address this fragility in the sparse view setting [62,34,21,24,9,6]. One line of research [62,6] aggregates available learning priors from adequate pre-training on large-scale datasets. Other approaches propose various regularizations on color and geometry of different views [9,21,24,34]. However, most aforementioned works still necessitate multiple view inputs, with a minimum requirement of three views [34,9].

In contrast to previous works, we push the setting of sparse views to the extreme by training a neural radiance field on only one single view. To our best knowledge, few efforts have been made to explore this circumstance before. PixelNeRF [62] takes the first attempt by pre-training a feature extractor on a large-scale dataset. Although they report impressive results on simple objects (e.g., ShapeNet dataset [5]), their performance on complex scenes [22] is less than satisfactory. Others [25,41] demonstrate good performance on novel-view synthesis. However, their platforms are based on other techniques (e.g., multiplane images). Different from those previous research, our work aims at training the neural radiance field from scratch, without bells and whistles, to generate photo-realistic novel views of complex scenes.

Nevertheless, training a neural radiance field with a single image is frustratingly challenging. First and foremost, reconstructing an accurate 3D shape from a single image meets several hurdles. Previous research has addressed reconstructing different types of objects from a single image [56,37]. Especially, Pixel2Mesh [56] proposes to reconstruct the 3D shape from a single image and expressed it in a triangular mesh. PIFu [37] adopts a 3D occupancy field to recover high-resolution surfaces of humans. NID [57] utilizes a pre-trained dictionary to acquire implicit fields from sparse measurements. However, all these approaches count on the prior knowledge specific to a certain object class or instance. Thus it can not work for complex scene reconstruction. Moreover, even in the simpler 2D cases, the exploration of training on single images is still gaining much interest as an open problem up to now [40,42,44,51]. SIREN [44] introduces a periodic activation for implicit functions to better fit a single image. SinGAN [40] and InGAN [42] propose to train generative adversarial networks (GANs) using a single image as a reference. Their models can generate visually-

pleasing results of images with similar content, but their results often boiled down to approximately replicating or re-composing the patches or textural patterns from the given images, and hence cannot serve the purpose of modeling sophisticated 3D view transformations.

Our inspiration draws from generating pseudo labels according to the available single view, which enables us to design a semi-supervised training strategy to constrain the learned radiance field. Specifically, we design two categories of pseudo labels to capture complementary hidden information. The first one focuses on the geometry of the radiance field, where we reproject depth information between reference view and unseen views through image warping [19], thus ensuring multi-view geometry consistency of our trained radiance field. The second one focuses on the semantic fidelity of the unseen views. We utilize a discriminator and a pre-trained Vision Transformer (ViT [10]) to constrain the unseen views: the former helps improve each unseen view's local textures, while the latter focuses on the perceptual quality of their global structures.

Our main contributions can be summarized as follows:

- We propose SinNeRF, a novel semi-supervised framework to train a neural radiance field in complex scenes effectively, using a single reference view.
- We introduce and propagate geometry and semantic pseudo labels to jointly guide the progressive training process. The former is inspired by image warping to ensure multi-view geometry consistency, and the latter enforces the perceptual quality of local textures as well as global structures.
- We conduct extensive experiments on complex scene benchmarks and show that SinNeRF can yield photo-realistic novel-view synthesis results without bells and whistles. Under the single image setting, SinNeRF significantly outperforms state-of-the-art NeRF baselines in all cases.

2 Related Works

2.1 Neural Radiance Field

Neural Radiance Fields (NeRFs) [31] have demonstrated encouraging progress for view synthesis by learning an implicit neural scene representation. Since its origin, tremendous efforts have been made to improve its quality [52,2,3,17,46,7], speed [32,36,47,15], artistic effects [53,13,20], and generalization ability [6,58,29,62]. Specifically, Barron *et al.* [2] propose to cast a conical frustum instead of a single ray for the purpose of anti-aliasing. Mip-NeRF 360 [3] further extends it to the unbounded scenes with efficient parameterization. KiloNeRF [36] speeds up NeRF by adopting thousands of tiny MLPs. MVSNeRF [6] extracts a 3D cost volume [60,16] and renders high-quality images from novel viewpoints on unseen scenes. The most related works to SinNeRF target the sparse view setting [62,9,34,21] Especially, DS-NeRF [9] adopts additional depth supervision to improve the reconstruction quality. RegNeRF [34] proposes a normalizing flow and depth smoothness regularization. DietNeRF [21] utilizes the CLIP embeddings [35] to add semantic constraints for unseen views. However, the CLIP

embeddings can only be obtained from low-resolution inputs due to memory issues. Thus it struggles to obtain texture details. Meanwhile, these methods can only perform well on at least two or three input views. PixelNeRF [62] utilizes a ConvNets encoder to extract context information by large-scale pre-training, and successfully renders novel views from a single input. However, it can only work on simple objects (e.g., ShapeNet [5]) while the results on complex scenes remain unknown. In our work, we focus on the challenging setting of using only one single view without any pre-training on multi-view datasets.

2.2 Single View 3D Reconstruction

Single view 3D reconstruction is a long-standing problem. Early methods use shape-from-shading [11] or adopt texture [26] and defocus [14] cues. These techniques rely on the existing regions of the images using a depth cue. More recent approaches hallucinate the invisible parts using learned priors. Johnston *et al.* [23] adopt an inverse discrete cosine transform decoder. Fan *et al.* [12] directly regresses the point clouds. Wu *et al.* [59] learns a mapping from input images to 2.5D sketches and maps the intermediate representations to the final 3D shapes. However, very few datasets are available for 3D annotation, and most of these methods use ShapeNet [5] which contains objects of simple shapes. There are also attempts to reconstruct the 3D shape of specific objects (e.g. humans). PiFU [37] utilizes a 3D occupancy field to recover the 3D geometry of clothed humans. DeepHuman [65] adopts an image-guided volume-to-volume translation framework. NormalGAN [55] conditions a generative adversarial network on the normal maps of the reference view.

Another line of research focuses on learning a 3D representation for view synthesis. Explicit representations involve volumetric representations [39,18,45], layer depth images (LDI) [49,41], and multiplane images (MPI) [30]. Implicit representations use coordinate-based networks to train a neural scene representation on one single view. PixelNeRF [62] takes the first attempt by utilizing a pre-trained feature extractor on large-scale dataset. Their results on complex scenes are less than satisfactory compared to their impressive results on simple objects from ShapeNet [5]. GRF [48] proposes a generative radiance field modeling 3D geometries by projecting the features of 2D images to 3D points. MINE [25] learns a continuous depth MPI and uses volumetric rendering to synthesize novel views. Our work is fundamentally different from existing works in these ways: 1) we train a neural scene representation from scratch without relying on pre-trained feature extractors or multi-plane images; 2) we conduct experiments on complex 3D environments and yield photo-realistic rendered results.

2.3 Single Image Training

Single image training is a field of great interest in 2D computer vision. SinGAN [40] and InGAN [42] propose a generative adversarial network trained using a single image as reference. Their models can generate visually-pleasing

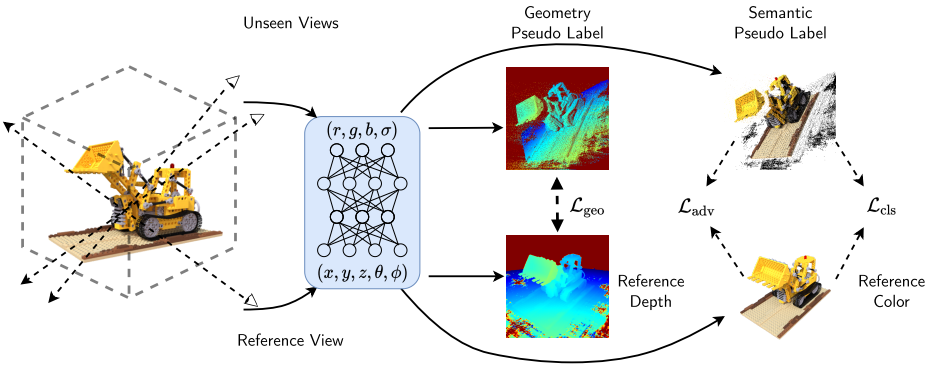

Fig. 2: An overview of our SinNeRF, where we synthesize patches from the reference view and unseen views. We train this semi-supervised framework via ground truth color and depth labels of the reference view and pseudo labels on unseen views. We use image warping to obtain geometry pseudo labels and utilize adversarial training as well as a pre-trained ViT for semantic pseudo labels.

results containing similar content of the image, but the diversity is limited, and their results often copy-paste different patches from the original image. Dmitry *et al.* [51] investigate the deep image prior of convolutional networks and show excellent results in image restoration. More recently, SIREN [44] proposes a periodic activation for implicit functions to fit a single image by supervising the gradients of networks. In this work, we make further attempts to adversarially train a radiance field using a single image.

3 Method

3.1 Overview

The setting of only one single view available is challenging for NeRF, as training directly on the available view leads to overfitting on the reference view and results in a collapsed neural radiance field. To tackle this problem, we build our SinNeRF as a semi-supervised framework to provide necessary constraints on unseen views. We treat the reference view with RGB and available depth as the labeled set, while the unseen views are considered as the unlabeled set. To help the neural radiance field render reasonable results on the unseen views, we introduce two types of supervision signals from the perspective of geometry and semantic constraints. We will first introduce the preliminary of neural radiance field and semi-supervised learning framework, then the progressive training strategies.

3.2 Preliminary

Neural Radiance Fields (NeRFs) [31] synthesize images sampling 5D coordinates (location (x, y, z) and viewing direction (θ, ϕ)) along camera rays, map them to

color (r, g, b) and volume density σ. Mildenhall *et al.* [31] first propose to use coordinate-based multi-layer perception networks (MLPs) to parameterize this function and then use volumetric rendering techniques to alpha composite the values at each location and obtain the final rendered images.

Given a pixel $r(t) = o + td$, where o is the camera origin and d is the ray direction, pixel's predicted color is defined as follows:

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

where $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s))ds\right)$, $\sigma(\cdot)$ and $c(\cdot, \cdot)$ are densities and color predictions from the network. Due to the computational cost, the continuous integral is numerically estimated using quadrature [31]. NeRF [31] optimize the radiance field by minimizing the mean squared error between rendered color and the ground truth color,

$$\mathcal{L}_{\text{pix}} = \sum_{r \in R_i} ||(C(r) - \hat{C}(r))||^2, \tag{2}$$

where R_i is the set of input rays during training.

3.3 Geometry Pseudo Label

Directly overfitting on the reference images leads to a corrupted neural radiance field collapsing towards the provided views. The issue is much more severe when there is only one training image. Without multi-view supervision, NeRF is not able to learn the inherent geometry of the scene and thus fails to build a view-consistent representation. Similar to previous works [61] to reconstruct a 3D shape from a single image, we start by adopting the depth prior to reconstructing reasonable 3D geometry. As suggested by [9], adding another depth supervision can significantly improve the learned geometry. However, since only a single training view is available in our setting, simply adopting depth supervision can not produce a reasonable 3D shape, as shown in Fig. 3.

To best utilize the available information in the reference view, we propose to propagate it to other views through image warping [19]. For pixel $p_i(x_i, y_i)$ in reference view I_{ref}, the corresponding pixel $p_j(x_j, y_j)$ in the j-th unseen view I_{unseen} can be formulated as:

$$p_j = K_{\text{unseen}} T(K_{\text{ref}}^{-1} Z_i p_i), \tag{3}$$

where Z_i is the available depth of reference view, T refers to the relationship between camera extrinsic matrices from I_{ref} to I_{unseen}, and K_{ref} and K_{unseen} refer to the camera intrinsic matrices. We further adopt the Painter's Algorithm [33]when multiple points in the reference view are projected to the same point in the unseen view and select the point with the smallest depth as the warping result.

Through image warping, we then obtain the depth map of an unseen view, which further serves as the pseudo ground truth label. Nevertheless, there is still an unavoidable gap between this pseudo ground truth and its real correspondence, since small misalignment in the predicted depth map can cause large errors when projected to other views. Moreover, it is quite common that the projected results contain some uncertain regions due to the occlusion. To regularize the uncertain regions in the warped results, we utilize the self-supervised inverse depth smoothness loss [54], which uses the second-order gradients of the RGB pixel value to encourage the smoothness of the predicted depths:

$$\mathcal{L}_{\text{smooth}}\left(d_i\right) = e^{-\nabla^2 \mathcal{I}(\mathbf{x}_i)}\left(\left|\partial_{xx} d_i\right| + \left|\partial_{xy} d_i\right| + \left|\partial_{yy} d_i\right|\right), \tag{4}$$

where d_i is the depth map, $\nabla^2 \mathcal{I}(\mathbf{x}_i)$ refers to the Laplacian of pixel value at location x_i. Similar to [54], we calculate this loss on a downscaled resolution.

We also reproject the unseen views back to the reference view to enforce geometry consistency. In summary, the geometry pseudo label is utilized as follows,

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_1(d_1, f(d_2)) + \mathcal{L}_1(f(d_1), d_2) + \lambda_4 \mathcal{L}_{\text{smooth}}, \tag{5}$$

where λ_4 is empirically set to be 0.1 in all our experiments, d_1 and d_2 refer to the depths of two views, and $f(\cdot)$ refers to the image warping result of the other view using the current view's depth information.

3.4 Semantic Pseudo Label

Since the rendered color and texture might still be inconstant across different views, image warping can only project depth information. We propose to adopt semantic pseudo labels to regularize the learned appearance representation. Unlike the geometry pseudo labels, where we enforce the consistency in 3D space, semantic pseudo labels are adopted to regularize the 2D image fidelity. Concretely speaking, we introduce a local texture guidance loss implemented by adversarial learning, and a global structure prior supported by a pre-trained ViT network. The two complementary guidances collaboratively help SinNeRF render visually-pleasing results in each view.

Local Texture Guidance The local texture guidance is implemented via a patch discriminator. The outputs from NeRF are considered as fake samples, and the patches randomly cropped from the reference view are regarded as real samples. Since the available training data are too limited, the discriminator tends to memorize the entire training set. To overcome this issue, we adopt differentiable augmentation [64] for our discriminator to improve its data efficiency:

$$\begin{aligned}
\mathcal{L}_{\text{D}} &= \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}\left[f_D(-D(T(\boldsymbol{x})))\right] + \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_D(D(T(G(\boldsymbol{z}))))\right], \\
\mathcal{L}_{\text{G}} &= \mathbb{E}_{\boldsymbol{z} \sim p(\boldsymbol{z})}\left[f_G(-D(T(G(\boldsymbol{z}))))\right], \\
\mathcal{L}_{\text{adv}} &= \mathcal{L}_{\text{D}} + \mathcal{L}_{\text{G}},
\end{aligned} \tag{6}$$

where T refers to the augmentation applied on both real and fake samples. We train the GAN framework using Hinge loss [27], so $f_D(x) = \max(0, 1 + x)$ and $f_G(x) = x$. The architecture of our discriminator is a cascade of convolutional layers. More details about the discriminator design is provided in the supplementary materials.

Global Structure Prior Vision transformers (ViT) have been proven to be an expressive semantic prior, even between images with misalignment [50,1]. Similar to [21], we propose to adopt a pre-trained ViT for global structure guidance, which enforces semantic consistency between unseen views and the reference view. Although there exists pixel-wise misalignment between the views, we observe that the extracted representation of ViT is robust to this misalignment and provides supervision at the semantic level. Intuitively, this is because the content and style of the two views are similar, and a deep network is capable of learning invariant representation.

Here we adopt DINO-ViT [4], a self-supervised vision transformer trained on ImageNet [8] dataset. Unlike DietNeRF [21] which utilizes a CLIP-ViT [35] and adopts its projected images embeddings as features, we directly extract the [CLS] token from DINO-ViT's output. This approach is more straightforward since the [CLS] token serves as a representation of an entire image [10]. The intuition also aligns with the recent findings of [50], where ViT architecture can capture semantic appearance after self-supervised pre-training. We calculate L_2 distance between the extracted features,

$$\mathcal{L}_{\text{cls}} = ||f_{\text{vit}}(A) - f_{\text{vit}}(B)||^2 \tag{7}$$

where $f_{\text{vit}}(\cdot)$ refers to the extracted [CLS] tokens. A and B are patches from the reference view and an unseen view, respectively.

3.5 Progressive Training Strategy

To stabilize the training of the GAN framework, we apply a progressive sampling strategy to the training of a single view neural radiance field.

Progressive Strided Ray Sampling: We start from utilizing a stride sampling [38] of ray generation and progressively reduce the stride size during training. This design enables our SinNeRF to cover a much larger region with a limited amount of rays. Specifically, the $K \times K$ patch P of stride s containing point (u, v) is defined as a set of 2D image coordinates,

$$\mathcal{P}(u, v, s) = \{(u + sx, v + sy) \mid x, y \in \{0, \ldots, K\}\}. \tag{8}$$

Under this circumstance, the NeRF is able to generate a $K \times K$ patch representing a large aspect of the scene. During training, we randomly sample two patches in each iteration, with the first one from the reference view and the other one from a random unseen view. After that, the collaborative local texture guidance

and global structure prior loss are applied on these patches to provide semantic guidance on the unseen views. Meanwhile, we obtain the geometry pseudo labels via image warping and add regularization on each patch's intersection with the corresponding patch's warped result. As the training goes into the latter stages, the stride s decreases so that the framework starts to focus on more local regions. Note that we randomly initialize the discriminator after reducing the stride size. This helps the discriminator focus on a fixed resolution, making the training more stable.

Progressive Gaussian Pose Sampling: After that, we propose to progressively enlarge the viewing angle during training. During training, we start at a local neighbor of the reference view and progressively rotate the camera pose more as the training proceeds. This helps the network to focus on dealing with the confident regions and stabilize training as the output image patches will have a good quality when the camera pose is only slightly different from the reference view. Specifically, we represent the distance between an unseen view and the reference view as Euler angles. Let (α, β, ϕ) denote the signed angles between the axis in the reference view's camera coordinate and the axis in the unseen views' camera coordinates. In each iteration, we sample α, β, ϕ each based on a Gaussian distribution $\mathcal{N}(0, \omega^2)$, where ω increases with more iterations.

We show the overall loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pix}} + \lambda_1 \mathcal{L}_{\text{geo}} + \lambda_2 \mathcal{L}_{\text{adv}} + \lambda_3 \mathcal{L}_{\text{cls}}, \tag{9}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weighting factors. We anneal the loss weight during training. In the early stages where we use a large stride and the patch covers the major regions of the original image, the global structure prior is given a large weight λ_3 compared to the weight of local texture guidance λ_2. As the training proceeds, we reduce the stride to focus on reconstructing the high-frequency details. Consequently, we reduce the weight of global structure prior λ_3 and increase the weight of local texture guidance λ_2. In all our experiments, λ_1, λ_2, and λ_3 are initialized to be 8, 0.1, and 0, respectively. During the training process, we gradually decrease λ_2 to 0 and increase λ_3 to 0.1 with a linear function.

4 Experiment

4.1 Implementation Details

We use the same architecture as the original NeRF paper [31]. During training iterations, we randomly sample two patches of rays from both the reference view and a random sampled unseen view. The size of patches on NeRF synthetic (Blender) dataset, Local Light Field Fusion (LLFF) dataset, and DTU dataset are set as 64×64, 84×63, and 70×56, respectively. The rendered patches are then sent to the discriminator and DINO-ViT network, where we additionally resize its input patches to 224×224 resolution to fit the input resolution of DINO-ViT architecture. We train our framework using RAdam optimizer [28],

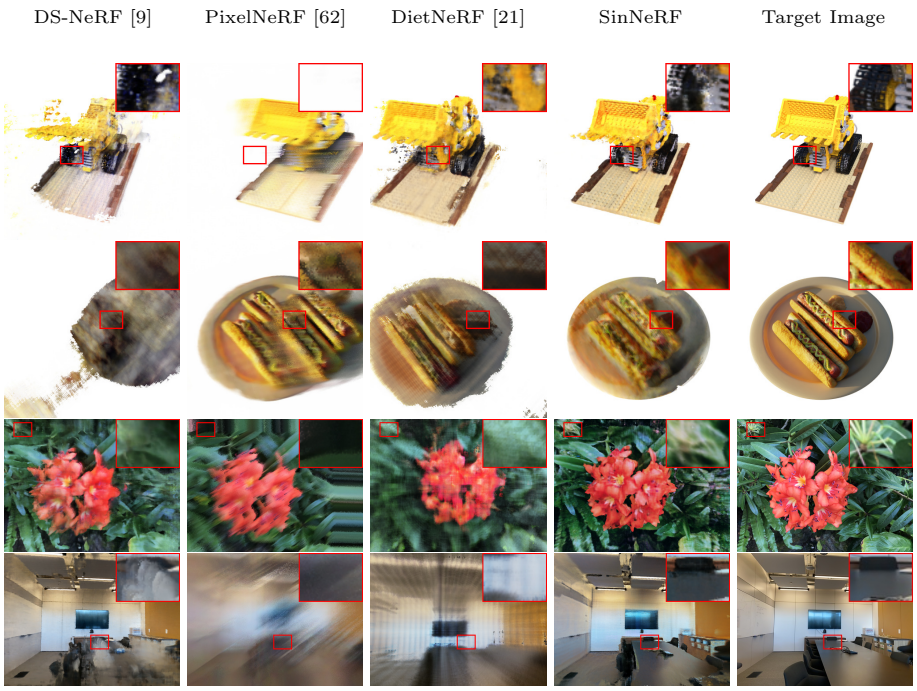

Fig. 3: Novel view synthesis results of different methods on NeRF synthetic and LLFF Dataset.

with an initial learning rate of $1e-3$. We decay the learning rate by half after every 10k iterations. The learning rate of the discriminator is kept to be 20% of the MLP's learning rate. The stride for sampling the patches starts at 6 and gradually reduces by 2 after every 10k iterations. All experiments of SinNeRF are conducted on an NVIDIA RTX A6000 GPU. The whole training process takes several hours for each scene. More implementation details and visual results are provided in the supplementary.

4.2 Evaluation Protocol

We perform experiments on NeRF synthetic dataset [31], Local Light Field Fusion(LLFF) dataset [30], and DTU dataset [22]. NeRF synthetic dataset contains complex objects with 360° view. LLFF provides complex forward-facing scenes. DTU consists of various objects placed on a table. We report metrics including PSNR, structural similarity index (SSIM), and LPIPS perceptual metric [63]. We compare our method with the state-of-the-art neural radiance field methods DietNeRF [21], PixelNeRF [62], and DS-NeRF [9]. We train DietNeRF and DS-NeRF for each scene since they are test-time optimization methods. As for PixelNeRF, we fine-tune the model on each scene before evaluation for a fair comparison.

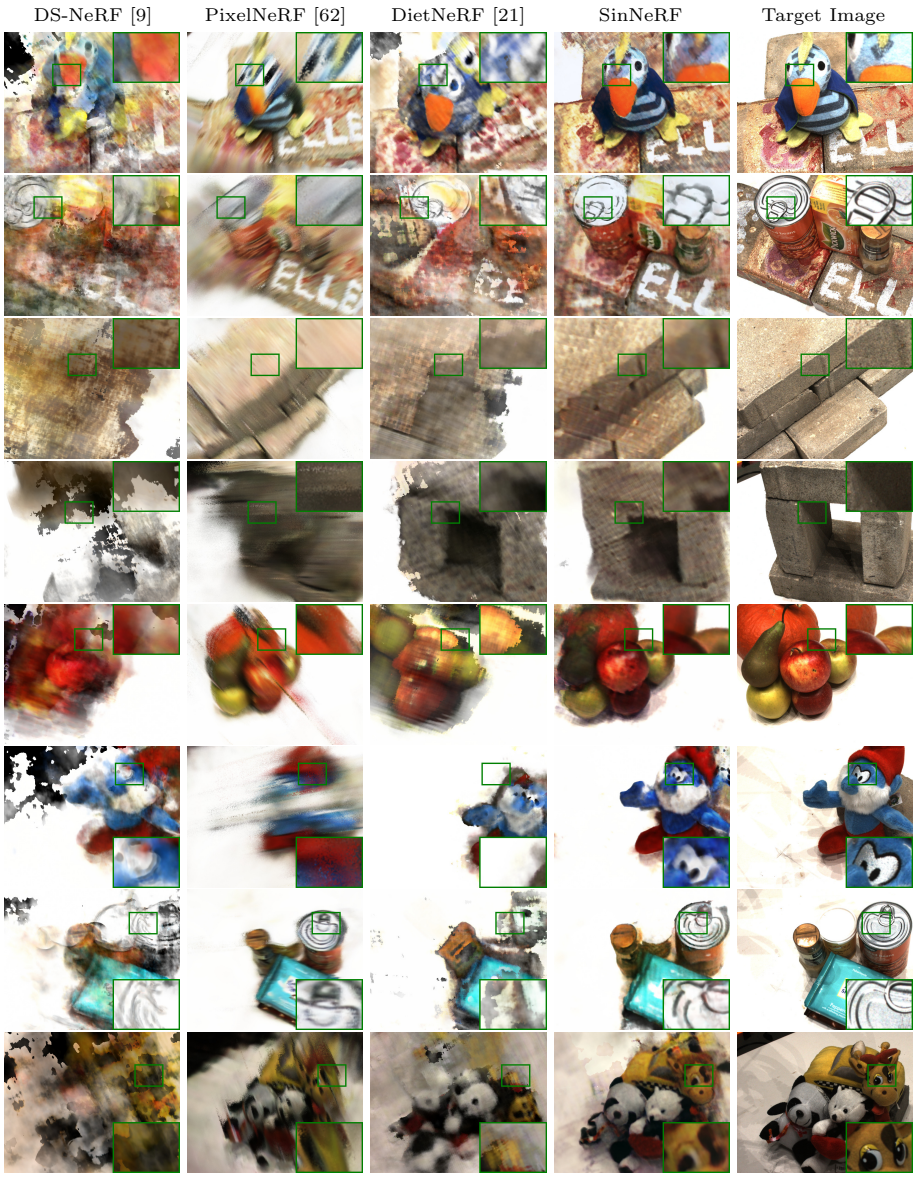

Fig. 4: Novel view synthesis results of different methods on DTU dataset.

4.3 View synthesis on NeRF Synthetic Dataset

For NeRF synthetic dataset, each scene is rendered via Blender. Both ground truth rendered images of 100 camera poses and the original blender files are provided. We randomly select a single view as the reference view and refer to its surrounding views as unseen views. Then we use blender to render the ground

	PSNR↑				SSIM↑				LPIPS↓			
	Lego	Hotdog	Flower	Room	Lego	Hotdog	Flower	Room	Lego	Hotdog	Flower	Room
DS-NeRF	16.62	14.16	16.92	17.44	0.77	0.67	0.41	0.65	0.1682	0.2956	0.3900	0.3986
DietNeRF	15.07	16.28	13.35	15.77	0.72	0.69	0.20	0.49	0.2063	0.2633	0.7526	0.7512
PixelNeRF	14.25	16.67	13.20	12.88	0.72	0.71	0.19	0.41	0.2171	0.2381	0.6378	0.7633
SinNeRF	20.97	19.78	17.20	18.85	0.82	0.77	0.41	0.67	0.0932	0.1700	0.3724	0.3796

Table 1: Quantitative evaluation of our method against state-of-the-art methods on the NeRF synthetic dataset (Lego and Hotdog) and LLFF dataset (Flower and Room).

	PSNR↑	SSIM↑	LPIPS↓
DS-NeRF	12.17	0.41	0.6493
DietNeRF	12.84	0.44	0.6469
PixelNeRF	12.06	0.42	0.6471
SinNeRF	16.52	0.56	0.5250

Table 2: Quantitative evaluation of our method against state-of-the-art methods on DTU dataset. We report average values across scenes.

truth of the unseen views by rotating the world-to-camera matrix. Specifically, we generate 60 test set images by rotating the camera around the y-axis uniformly in $[-30°, 30°]$. The quantitative results are shown in Tab. 1. Our method achieves the best results both in pixel-wise error and perceptual quality.

We show the novel view synthesis results in the first two rows of Fig. 3. Each row corresponds to a fixed camera pose, and each column contains the results of a method. One can see that our method preserves the best geometry as well as perceptual quality. DS-NeRF's output contains a wrong geometry at the top of the lego. This is because DS-NeRF only utilizes supervision on the reference view and does not perform warping to other views. PixelNeRF's results contain "ghost" hotdogs since they do not explicitly regularize the geometry. Optimizing on unseen views, DietNeRF produces appealing results, but unfortunately with flaws in the novel view's geometry (e.g., the objects are no longer in the center). The results are also blurry since their CLIP embeddings are obtained at a low resolution.

4.4 View synthesis on LLFF Dataset

For the local light field dataset, the images and the SfM results from colmap are provided. We randomly select a single view as the reference view and use its surrounding views as unseen views during training. For quantitative evaluation, we render the other views in the dataset whose ground truth images are available. We provide visual results in the last two rows of Fig. 3 and quantitative results in Tab. 1. Our method generates the most visually-pleasing results, while other methods tend to render obscure estimations on novel views. DS-NeRF shows realistic geometry, but the rendered images are blurry. PixelNeRF and DietNeRF

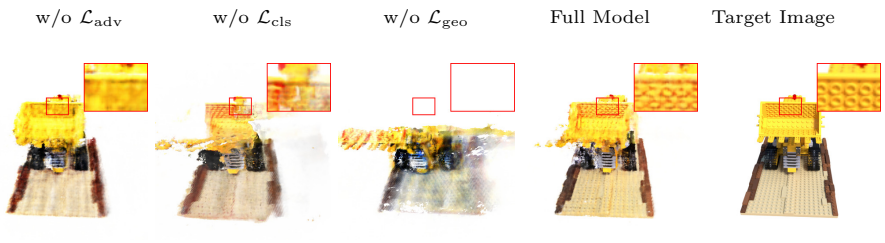

Fig. 5: Novel view synthesis from different variants of our proposed model.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o $\mathcal{L}_{geo}$	16.11 (-4.86)	0.74 (-0.08)	0.1919 (+0.0987)
w/o $\mathcal{L}_{cls}$	18.20 (-2.77)	0.76 (-0.06)	0.1348 (+0.0146)
w/o $\mathcal{L}_{adv}$	20.20 (-0.77)	0.79 (-0.03)	0.1306 (+0.0294)
Full Model	**20.97**	**0.82**	**0.0932**

Table 3: **Ablation study on variants of pseudo labels**. "w/o $\mathcal{L}_{adv}$" refers to the variant without the local texture guidance. "w/o $\mathcal{L}_{cls}$" refers to the variant without global structure prior. "w/o $\mathcal{L}_{geo}$" refers to removing the geometry pseudo labels and using depth supervision only on the reference view. Experiments are conducted on Lego scene.

present good structures but wrong geometry due to their lack of local texture guidance and geometry pseudo label.

4.5 View synthesis on DTU Dataset

For each scene in DTU dataset, 49 images and their fixed camera poses are provided. We use camera 2 as the reference view because its images contain most parts of the scene. We use 10 nearby cameras from the dataset as unseen views during training. Since the ground truth of these nearby views are provided, we render these views for quantitative evaluation. We provide visual results in Fig. 4 and quantitative results in Tab. 2. Our method demonstrates the most visually-pleasing results as well as the best quantitative performance. DS-NeRF generates realistic geometry, but the results contain severe artifacts. PixelNeRF and DietNeRF obtain a pleasing overall looking but suffer from wrong geometry.

4.6 Ablation Study

Variants of pseudo labels . In this section, we study the effectiveness of each component of our proposed method. We evaluate on the lego scene and provide the results in Fig. 5 and Tab. 3. Removing adversarial training leads to blurry artifacts. This is because the $\mathcal{L}_{cls}$ is only beneficial when the extracted patch has a receptive field large enough to cover the major structure of the image. The variant without global structure prior contains wrong structure in novel views,

	PSNR↑	SSIM↑	LPIPS↓
style	15.49 (-5.48)	0.73 (-0.09)	0.2046 (+0.1114)
self-similarity	18.67 (-2.30)	0.81 (-0.01)	0.1075 (+0.0143)
content	19.20 (-1.76)	0.80 (-0.02)	0.1138 (+0.0206)
$[CLS]$ **(ours)**	**20.97**	**0.82**	**0.0932**

Table 4: **Ablation study on different choices of the global structure prior.** Here "content loss" refers to calculating L_1 loss on the feature space of pretrained VGG-16 network [43]. "style loss" refers to minimizing gram matrix from the output of pre-trained VGG-16 network [43]. "self-similarity loss" [50] refers to calculating the self-similarity of the keys in ViT's self-attention layer. The $[CLS]$ denotes our proposed one, where we adopt the $[CLS]$ token from pre-trained DINO-ViT approaches. $\mathcal{L}_{\text{cls}}$. Experiments are conducted on Lego scene.

which is due to the missing guidance on the overall semantic structure. Although there are still geometry pseudo labels available, the projected depth information only provides partial guidance and leaves the occluded regions unconstrained. Finally, the variant without geometry pseudo labels suffers from wrong geometry. There is only depth supervision of the reference view, and the unseen views are not properly regularized.

Different choices of the global structure prior. We study different model choices for our global structure prior in this section. The global structure prior is designed to focus on the overall semantic consistency between the unseen views and the reference view regardless of the pixel misalignment. Following this direction, we evaluate different architectures including both the ConvNets and ViTs. As shown in Tab. 4, we evaluate different kinds of the global structure prior by conducting experiments on the "lego" scene, including adopting the content, style, and self-similarity losses from a pre-trained VGG network between unseen views and reference view or minimizing the distance between the outputs of $[CLS]$ token from DINO-ViT [4] architecture. The quantitative results demonstrate that DINO-ViT shows a stronger global structure prior, suggesting that it is more robust to pixel misalignment.

5 Conclusions

We present SinNeRF, a framework to train a neural radiance field on a single view from a complex scene. SinNeRF is based on a semi-supervised framework, where geometry pseudo label and semantic pseudo label are synthesized to stabilize the training process. Comprehensive experiments are conducted on complex scene datasets, including NeRF synthetic dataset, Local Light Field Fusion (LLFF) dataset, and DTU dataset, where SinNeRF outperforms the current state-of-the-art NeRF frameworks. However, similar to most NeRF approaches, one limitation of SinNeRF is the training efficiency issue, which could be one of our future directions to explore further.

References

1. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 (2021)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. arXiv preprint arXiv:2111.12077 (2021)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
6. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
7. Chen, T., Wang, P., Fan, Z., Wang, Z.: Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15191–15202 (2022)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. arXiv preprint arXiv:2107.02791 (2021)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Durou, J.D., Falcone, M., Sagona, M.: Numerical methods for shape-from-shading: A new survey with benchmarks. Computer Vision and Image Understanding **109**(1), 22–43 (2008)
12. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017)
13. Fan, Z., Jiang, Y., Wang, P., Gong, X., Xu, D., Wang, Z.: Unified implicit neural stylization. arXiv preprint arXiv:2204.01943 (2022)
14. Favaro, P., Soatto, S.: A geometric approach to shape from defocus. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(3), 406–417 (2005)
15. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
16. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2495–2504 (2020)

17. Guo, Y.C., Kang, D., Bao, L., He, Y., Zhang, S.H.: Nerfren: Neural radiance fields with reflections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18409–18418 (2022)
18. Henzler, P., Mitra, N.J., Ritschel, T.: Learning a neural 3d texture space from 2d exemplars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8356–8364 (2020)
19. Huang, B., Yi, H., Huang, C., He, Y., Liu, J., Liu, X.: Mˆ3vsnet: Unsupervised multi-metric multi-view stereo network. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 3163–3167. IEEE (2021)
20. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
21. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021)
22. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 406–413 (2014)
23. Johnston, A., Garg, R., Carneiro, G., Reid, I., van den Hengel, A.: Scaling cnns for high resolution volumetric reconstruction from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 939–948 (2017)
24. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. arXiv preprint arXiv:2112.15399 (2021)
25. Li, J., Feng, Z., She, Q., Ding, H., Wang, C., Lee, G.H.: Mine: Towards continuous depth mpi with nerf for novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12578–12588 (2021)
26. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2041–2050 (2018)
27. Lim, J.H., Ye, J.C.: Geometric gan. arXiv preprint arXiv:1705.02894 (2017)
28. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020) (April 2020)
29. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7824–7833 (2022)
30. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
31. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European conference on computer vision. pp. 405–421. Springer (2020)
32. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv preprint arXiv:2201.05989 (2022)
33. Newell, M.E., Newell, R., Sancha, T.L.: A solution to the hidden surface problem. In: Proceedings of the ACM annual conference-Volume 1. pp. 443–450 (1972)

34. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. arXiv preprint arXiv:2112.00724 (2021)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
36. Reiser, C., Peng, S., Liao, Y., Geiger, A.: Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14335–14345 (2021)
37. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2304–2314 (2019)
38. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. Advances in Neural Information Processing Systems **33**, 20154–20166 (2020)
39. Seitz, S.M., Dyer, C.R.: Photorealistic scene reconstruction by voxel coloring. International Journal of Computer Vision **35**(2), 151–173 (1999)
40. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4570–4580 (2019)
41. Shih, M.L., Su, S.Y., Kopf, J., Huang, J.B.: 3d photography using context-aware layered depth inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8028–8038 (2020)
42. Shocher, A., Bagon, S., Isola, P., Irani, M.: Ingan: Capturing and remapping the" dna" of a natural image. arXiv preprint arXiv:1812.00231 (2018)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
44. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems **33**, 7462–7473 (2020)
45. Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., Zollhofer, M.: Deepvoxels: Learning persistent 3d feature embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2437–2446 (2019)
46. Suhail, M., Esteves, C., Sigal, L., Makadia, A.: Light field neural rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8269–8279 (2022)
47. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022)
48. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15182–15192 (2021)
49. Tulsiani, S., Tucker, R., Snavely, N.: Layer-structured 3d scene inference via view synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 302–317 (2018)
50. Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.: Splicing vit features for semantic appearance transfer. arXiv preprint arXiv:2201.00424 (2022)

51. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9446–9454 (2018)
52. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. arXiv preprint arXiv:2112.03907 (2021)
53. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3835–3844 (2022)
54. Wang, C., Buenaposada, J.M., Zhu, R., Lucey, S.: Learning depth from monocular videos using direct methods. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2022–2030 (2018)
55. Wang, L., Zhao, X., Yu, T., Wang, S., Liu, Y.: Normalgan: Learning detailed 3d human from a single rgb-d image. In: European Conference on Computer Vision. pp. 430–446. Springer (2020)
56. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–67 (2018)
57. Wang, P., Fan, Z., Chen, T., Wang, Z.: Neural implicit dictionary via mixture-of-expert training. In: International Conference on Machine Learning (2022)
58. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
59. Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., Tenenbaum, J.: Marrnet: 3d shape reconstruction via 2.5 d sketches. Advances in neural information processing systems **30** (2017)
60. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European conference on computer vision (ECCV). pp. 767–783 (2018)
61. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 204–213 (2021)
62. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
63. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
64. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
65. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7739–7749 (2019)