SPARF: Large-Scale Learning of 3D Sparse Radiance Fields from Few Input Images

Abdullah Hamdi [1,2,3] Bernard Ghanem [2] Matthias Nießner [1]
[1]Technical University of Munich (TUM) [2]King Abdullah University of Science and Technology (KAUST)
[3]Visual Geometry Group, University of Oxford
abdullah.hamdi@eng.ox.ac.uk

Abstract

Recent advances in Neural Radiance Fields (NeRFs) treat the problem of novel view synthesis as Sparse Radiance Field (SRF) optimization using sparse voxels for efficient and fast rendering [14, 43]. In order to leverage machine learning and adoption of SRFs as a 3D representation, we present SPARF, a large-scale ShapeNet-based synthetic dataset for novel view synthesis consisting of ∼ 17 million images rendered from nearly 40,000 shapes at high resolution (400 × 400 pixels). The dataset is orders of magnitude larger than existing synthetic datasets for novel view synthesis and includes more than one million 3D-optimized radiance fields with multiple voxel resolutions. Furthermore, we propose a novel pipeline (SuRFNet) that learns to generate sparse voxel radiance fields from only few views. This is done by using the densely collected SPARF dataset and 3D sparse convolutions. SuRFNet employs partial SRFs from few/one images and a specialized SRF loss to learn to generate high-quality sparse voxel radiance fields that can be rendered from novel views. Our approach achieves state-of-the-art results in the task of unconstrained novel view synthesis based on few views on ShapeNet as compared to recent baselines. The SPARF dataset is made public with the code and models on the project website abdullahamdi.com/sparf.

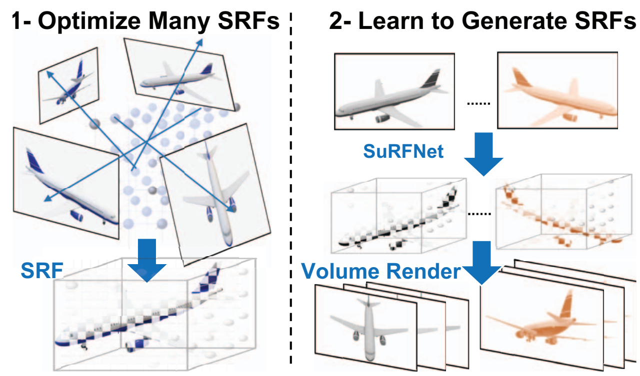

Figure 1. **Distribution of Radiance Fields.** We treat Sparse Radiance Field (SRF) as a 3D data structure and learn the conditional generation of SRFs from few input images for the task of novel view synthesis. In order to do this, we build SPARF, a large-scale dataset of SRFs.

1. Introduction

Although we observe the surrounding world only as a stream of 2D images, it is undeniably 3D. The goal of recovering this underlying 3D from 2D observations has been a longstanding goal of computer vision. The task of inverting the rendering process that creates the 2D projections we observe by trying to construct the 3D world is known as Vision as Inverse Graphics (VIG) [7, 28, 72, 26]. With the emergence of deep learning applications in computer graphics and the availability of 3D datasets, several approaches address the 3D generation task directly from 3D data, without relying on appearance [44, 50, 22, 1, 67, 29]. However, recent developments in differentiable rendering have refueled the VIG direction, which facilitates using gradients of the rendering process to optimize for the underlying 3D setup based on image observations [51, 58, 16, 73, 19, 20, 17, 39, 41, 34, 37, 55, 33]. More specifically, Neural Radiance Fields (NeRFs) [41, 48, 71, 6] show impressive performance on novel view synthesis by optimizing volumetric radiance fields on a large number of posed multi-view images.

Various subsequent works addressed NeRFs' shortcomings, as rendering speed [43, 14, 70], training size requirement [71, 6, 42], or pose requirements [48, 36]. The seminal work of Plenoxels [14] showed that the MLP network is not necessary for quick optimization and volumetric rendering of the radiance fields. However, these recent methods are still optimization-based, where a single scene representation is optimized without any generalization to new scenes/objects [43, 14]. In this work, we treat Sparse Radiance Fields (SRFs) as a 3D data structure and try to learn a generative model (dubbed *SuRFNet*) on the distribution of sparse-voxel radiance fields conditioned on a few images to generalize to unseen 3D shapes (see Figure 1).

Attribute	Posed Multi-View Datasets SRN [56]	DTU[24]	NMR [46]	RTMV [57]	SPARF (ours)
Number of Classes	2	N/A	13	N/A	13
Number of Scenes/Objects	3,511	124	43,756	2,000	39,705
Image Resolution	128	1,200	64	1,600	400
Number of Radiance Fields	0	0	0	2,000	**1,072,008**
Real/Synthetic	synthetic	real	synthetic	synthetic	synthetic
View Setup	sphere	random	circle	hemisphere	sphere
Total Number of Images	265,550	4,235	1,050,144	300,000	**17,073,150**
Views per Model	50	N/A	24	N/A	430
Dataset Size (GB)	5.8	1	33	2,520	3,432

Table 1. **Comparison of Different Posed Multi-View Datasets**. We compare posed multi-view datasets to our large-scale SPARF dataset.

In order to train deep learning models on 3D data to generalize to unseen examples, the dataset size should be in tens of thousands of samples [64, 5]. However, current posed multi-view datasets are not suitable for leveraging the power of deep networks as can be seen in Table 1. The image resolution is either too low (*e.g.* 64×64 in [46]), the samples don't fall into distinct shape categories with similar numbers of views [24, 52, 25], or lack diversity in the samples and classes [56]. For these reasons, we construct a large and high-resolution dataset (SPARF) of posed multi-view images from ShapeNet [5] that correspond to the same 13 classes originally used in the NMR dataset [46], but with an order of magnitude more images and pixels (17M *vs*. 1M images and 400×400 *vs*. 64×64 pixels). We also provide more than *one million* optimized sparse radiance fields of spherical harmonics and densities that allow for the novel view synthesis of the 40K models using Plenoxels [14].

The idea of learning a prior (2D CNN/ViT) on radiance fields in order to enhance the few-view setup of novel view synthesis is previously investigated by several works [71, 53, 32]. However, we propose SuRFNet to *directly* learn from the 3D sparse radiance fields, by optimizing *partial* SRFs from the few images and training a generalizable network that converts these partial SRFs to *whole* SRFs in a supervised fashion. Such a 3D setup benefits from grid-based 3D learning, creating a 3D prior that ensures multi-view consistency, especially when rendering from out-of-distribution views. Also, this 3D sparse voxel setup benefits from the advancements in fast volume rendering [43, 14], allowing for end-to-end deep learning pipelines that harness volume rendering. To the best of our knowledge, our SURFNet is the first model that learns to generate 3D radiance fields for unseen objects at test time with only a few/single views by learning from the distribution of radiance fields in 3D.

Contributions: (i) To facilitate the application of deep learning on radiance fields, we provide a new Posed Multi-view dataset (SPARF) that is an order of magnitude larger than others (around 40K 3D models). The dataset includes a total of one million optimized Sparse Radiance Fields (SRFs) with multiple voxel resolutions, which allows for high-quality novel view synthesis and will be made publicly available. **(ii)** We propose a novel architecture and a pipeline (SuRFNet) equipped with a specialized SRF-loss to generate voxel-based radiance fields from a few images based on learning to complete partial radiance fields. SuRFNet improves the performance of unconstrained novel view synthesis based on few views compared to state-of-the-art methods.

2. Related Work

Learning 3D Shapes. Several works aim to predict the geometry of 3D shapes given several input images, by directly optimizing the vertices of a template mesh through differentiable projections or through fitting a network [58, 16, 73, 17, 39, 15, 49]. Other works use MLPs as a deep prior to the optimized mesh [21, 62]. Alternately, some methods try to learn the distribution of 3D meshes by optimizing 3D generators independent of how the meshes look when rendered, solely based on the available 3D data and heuristic regularizers [44, 50, 9]. Point cloud methods offer an alternative to the mesh complex topology by learning generative models on the point clouds themselves, *e.g.* by using an Auto Encoder [1, 67] or a GAN framework [1, 29]. The implicit representation paradigm offers an alternative to meshes for smooth and detailed shape representation. These methods learn a continuous implicit representation of shapes by learning the Signed Distance Functions or occupancy of the object through MLPs [47, 38, 69, 45, 4, 3, 54, 34]. In this work, the scope focuses on the quality of the rendering from novel views and not on 3D reconstruction.

Neural Radiance Fields (NeRFs). NeRFs [41] proved to be a successful popularizing in implicit volume representation and novel view synthesis. They define an implicit field and learn an MLP that predicts the RGB and density value of that 3D field given a set of posed images. NeRFs shoot rays on the volume and integrate the predictions to obtain individual pixel values. This formulation, however,

Figure 2. **SPARF: a Large Dataset for 3D Shapes Radiance Fields and Novel Views Synthesis**.

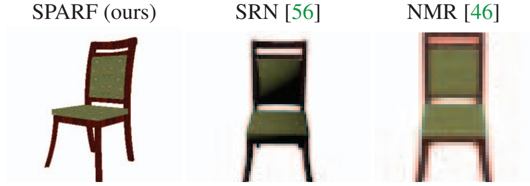

Figure 3. **SPARF *vs*. other Datasets**. SPARF offers a large-scale high-resolution dataset compared to other posed multi-view datasets. We show the same chair here on SPARF, SRN, and NMR (please zoom in for differences). This highlights the huge quality gap between SPARF and other ShapeNet-based datasets.

has many drawbacks including large memory and compute requirements, inability to model dynamic scenes, posed image requirements, and the limitation to small 3D objects or rooms [36, 48, 70, 45, 71, 30]. To address the speed limitation, PlenOctreeNeRF [70] stores the precomputed RGB, density, and spherical harmonics in the 3D volume as an Octree data structure for fast inference. Plenoxels [14] optimize the density and spherical harmonics on sparse voxels with a TV loss and perform ray marching for rendering from novel views. Similarly, INGP [43] uses multi-resolution voxel hashing to perform a real-time rendering of radiance fields, demonstrating that the redundancy of the MLP in NeRFs. We build on these observations and build the SPARF dataset of sparse voxel radiance fields in order to facilitate learning on these SRFs as 3D data structures instead of as just side outcomes of a volumetric optimization.

Few-Image NeRFs. To address the original NeRF's requirement of many posed images, several methods were proposed. The seminal work PixelNerf [71] is the first to reduce the image data requirements in order to learn a NeRF by using a trained CNN prior that can allow for transferable representation between scenes. Similarly, MVSNerf [6, 18], AutoRF [42], and ShaRF [53] learn a CNN prior to generalize across scenes. IBRNet [59] learns to render novel views based on neighboring views and optimized neural volume representation. More recently, VisionNerf [32] proposes to use a ViT [12] to extract global features from the input images to enhance the capability of the NeRF MLP to predict the radiance field when one image is used as input. Unlike these works, we propose SuRFNet to directly learn from the 3D sparse radiance fields. Such a 3D setup benefits from structured 3D learning, creating a 3D prior that guarantees multi-view consistency while benefiting from the speed of recent voxel-based methods. A concurrent work by Guo *et al*. [18] learns a 3D prior based on a perceptual loss, but does not use 3D supervision and only uses dense voxels (limiting the pipeline to the low resolution of 64^3).

Datasets for novel view synthesis. Several datasets were proposed to support the task of novel view synthesis. NeRF [41] introduced 8 synthetic scenes with 360-degree views. Wang *et al*. [60] introduced Google Scanned Objects. Other datasets for training multi-view algorithms include DTU [24], LLFF [40], Tanks and Temples [27], Spaces [13], RealEstate10K [75], SRN [56], Transparent Objects [23], ROBI [66], CO3D [52], SAPIEN [65], and BlendedMVS [68]. Recently, RTMV [57] introduced a ray-traced posed multi-view dataset with 2000 scenes and high-resolution images. Unfortunately, the current posed multi-view datasets commonly used in NeRF research are either small in image resolution or the number of posed images (SRN [56] and NMR [46]), small in the number of scenes/shapes and classes (Synthetic NeRFs [41]), or lack structure (DTU [24], RTMV [57]) and Objeverse [11]. A detailed multi-attribute comparison is provided in Table 1.

3. SPARF: a Large Dataset of 3D Shapes Radiance Fields

3.1. Motivations for SPARF

One of the goals of this work is to learn to generate high-quality SRFs in one forward pass of a deep network to enable fast novel view synthesis. In order to do this, harnessing the power of deep networks would require a large dataset of SRFs that follow a distinct distribution (*e.g*. different class categories). Datasets similar to SPARF (SRN [56], NMR [46]) form the basis of many advancements that pushed the community of neural radiance fields forward [71, 32]. As can be seen in Table 1 and Figures 2 and 3, SPARF naturally extends both datasets in orders of magnitude in resolution, scale, and diversity, benefiting the entire research community. Furthermore, the density of views ($\sim$ 17M) is higher than in previous datasets. This setup helps in studying out-of-distribution generalization of the views as a new benchmark that was neglected in previous works (see Section 5.3).

3.2. Dense Posed Multi-view Image Dataset

The first step in collecting the desired large high-quality radiance field dataset is to collect a synthetic posed multi-view dataset. We use ShapeNet Core 55 [5] as the data of choice for 3D shapes. For rendering, we used Pyglet [2] API through Trimesh library [10]. The renderer is based on OpenGL [63] rasterizer to render over 17 million images of

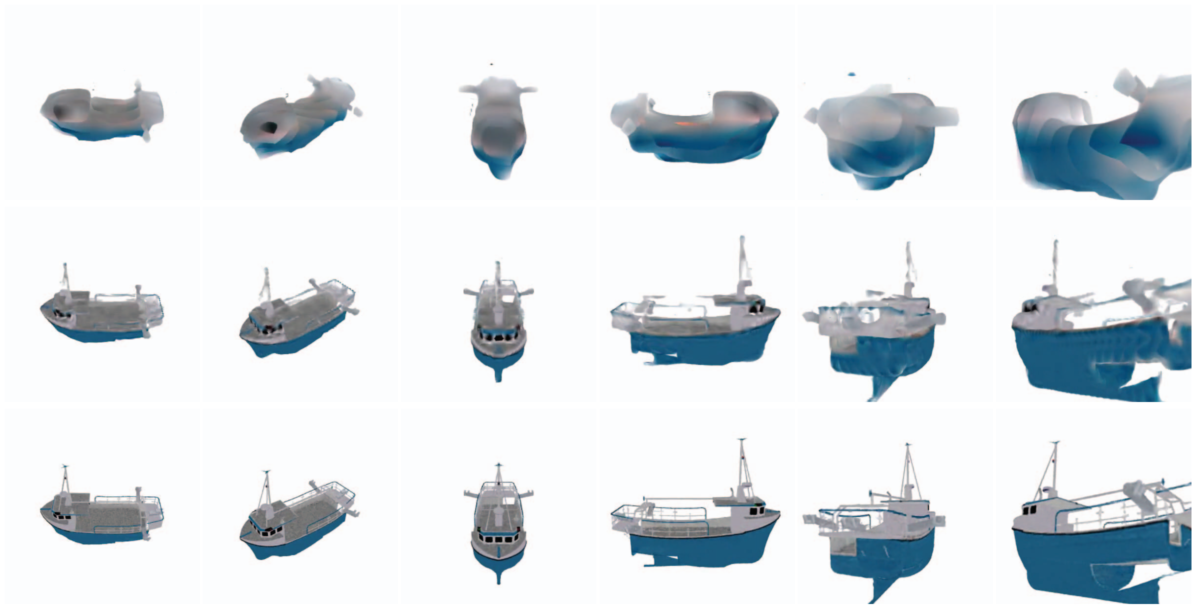

Figure 4. **SRFs: The optimized Sparse Radiance Fields in SPARF**. A total of one million SRFs have been collected in SPARF, including on multiple voxel resolutions: 32 (*top*), 128 (*middle*), and 512 (*bottom*).

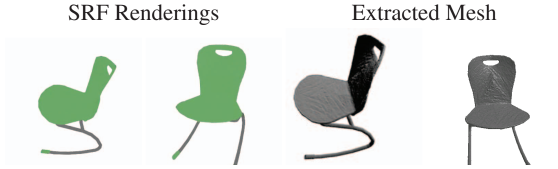

Figure 5. **Extracting 3D Meshes from SRFs**. Since SPARF and SuRFNet lie on the 3D voxel's space, extracting the mesh is straightforward with one pass of MarchingCubes [35].

around 40,000 shapes from 13 different classes at a high resolution of 400×400. Every shape is rendered equidistantly from 400 views distributed in a spherical configuration surrounding the object, including from the bottom (see Figure 2 for examples). An additional 20 views are rendered from random views from the same distance as test views for novel view synthesis tasks. Furthermore, an additional 10 views are rendered randomly from random distances bounded by a reasonable range, such that at least a part of the object is guaranteed to be visible. This last set is aimed at robustness purposes to test whether novel view synthesis methods can generalize to out-of-distribution posed views.

3.3. Multi-Resolution 3D SRFs

Sparse Radiance Field (SRF) can be defined as a voxel grid of dimension $1 + d$, where d is the dimension of radiance colors $\boldsymbol{\rho}_{i,j,k} \in \mathbb{R}^d$ at that specific (i, j, k) indexed voxel in addition to one dimension for density $\alpha_{i,j,k} \in \mathbb{R}$. We assume that the grid is of size H in each of the three dimensions: $\mathcal{X} \in \mathbb{R}^{H^3 \times (1+d)}$. Since the SRF is sparse, it can be represented with the COO format [8] as a set of M tuples of positive integer coordinates $\mathbf{c} \in \mathbb{Z}^+$ and features $\mathbf{f} \in \mathbb{R}^{d+1}$ with the sparsity of $1 - \frac{M}{H^3}$ as follows:

$$\mathcal{X}_{\text{non-empty}} = \{(\mathbf{c_m}, \mathbf{f_m})\}_{m=1}^{M} \tag{1}$$

The ordering of the set of tuples is arbitrary, but the features $\mathbf{f}_m$ consist of the density $\alpha_{i,j,k}$ and radiance colors $\boldsymbol{\rho}_{i,j,k}$ at that location $\mathbf{c}_m = (i, j, k)$. For the radiance field colors, we use the spherical harmonics proposed in Plenoxels [14] for view-dependent learning of radiance common in NeRFs [41]. The SRF can be viewed as a encoding of the NeRF MLP into sparse voxels for efficient optimization and volume rendering. In many 3D object tasks, a coarse-to-fine approach is followed [31], demanding multiple resolutions. Hence, we collect the SPARF with multiple resolutions $H \in \{32, 128, 512\}$, as shown in Figure 4. We used an adaptation of Plenoxels [14] to collect the dataset of a total of one million SRFs as we detail next. In order to scale up the Plenoxels optimization for this huge number of shapes and variants, we utilize a large number of images in the SPARF dataset to reduce the iterations to a minimum number while maintaining a high average PSNR across the dataset for the collected SRFs across the multiple resolutions. highly detailed 3D meshes can be extracted easily from the collected SRFs as can be seen in Figure 5.

3.4. Representing Images with Partial SRFs

In addition to collecting the "whole" part of SPARF that utilized all 400 images for every shape in optimizing the SRFs, we collect "partial" SRFs. These are SRFs that are quickly optimized on only a small number of images (1 or 3) randomly sampled from all 400 images, resulting in multiple partial SRF variants of that shape (see Figure 6). The Million SRFs dataset consists of two types of SRFs, either optimized

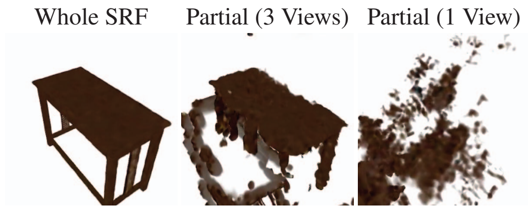

Figure 6. **Whole *vs*. Partial SRFs**. The partial SRFs are used instead of the few images that generated them as input to the learning pipeline to generate the whole SRFs

using the entire set of 400 views to generate one Plenoxel [14] per object, or optimized using multiple partial views from each object to generate multiple Plenoxel variants (see Figure 7). The partial SRFs were utilized as input to our SuRFNet pipeline (refer to Figure 8). The release of these partial SRFs should assist in 3D pipelines that consider radiance fields as a 3D data structure [18], as we demonstrate later in Section 6.1.

4. Learning to Generate SRFs

4.1. SuRFNet: Sparse Radiance Fields Network

3D Pipeline. Previous methods in few-views novel view synthesis (*e.g.* PixelNerf [71] and VisionNerf [32]) learn 2D priors to generalize across different shapes. On the other hand, we propose *SuRFNet* to distill the image views into partial 3D SRFs and then perform the learning in the 3D sparse voxel space to generate full SRFs based on the partial SRFs. Such a 3D Learning setup ensures multi-view consistency, especially when rendering from out-of-distribution views (as we show in Section 5.3). The input to the pipeline is the input partial SRFs from Section 3.4, where the goal is learning a generalizable network that converts partial SRFs to whole SRFs as can be seen in Figure 8. We leverage the Minkowski Net [8] as the 3D sparse convolution network of choice. However, As this novel setup of learning SRFs poses new challenges, we propose several modifications to the typical pipeline of training MinkoskiNet (the *type* of losses and *where* to define them).

Challenges of Learning SRFs. As a 3D data structure, SRF is an irregular volumetric representation that does not necessarily reflect the underlying 3D shape/scene, as it results from the optimization of posed images into volume. Many of the non-empty voxels have low densities and do not affect the volume rendering, but include color information that can confuse the network. Also, small errors in predicting the densities or radiance colors can result in large distortions in the rendered images, hurting the overall performance of novel view synthesis. Another challenge in learning SRFs is vanishing gradients. The typical sparsity in our setup is $\sim 99\%$, and misalignment between the input SRF coordinates and output SRF coordinates can further harm the gradients and affect the learning process. In order to tackle these issues,

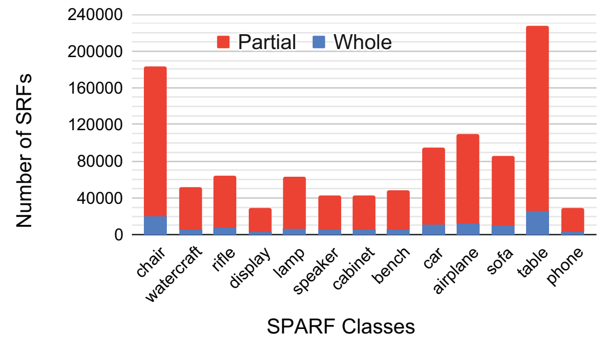

Figure 7. **SPARF Distribution**. We show the distribution of classes in SPARF and how the one million partial and whole SRFs are distributed. The numbers are equally distributed on three voxel resolutions: 512,128. and 32.

we propose three specialized losses detailed next.

4.2. SRF-Loss

Density loss. The goal of the density loss is to create a dense surface. We propose the following binary cross-entropy loss on the predicted densities α as follows:

$$\begin{aligned} L_{\alpha}\left(\mathcal{X}\,,\,\hat{\mathcal{X}}\right) &= -(\hat{\mathbf{y}}\log(\mathbf{y}) + (\mathbf{1}-\hat{\mathbf{y}})\log(\mathbf{1}-\mathbf{y})) \\ \text{s. t.}\quad \hat{\mathbf{y}} = \mathbb{1}\left(\mathcal{S}(\hat{\mathcal{X}}_{\alpha}) > \alpha_{\text{dense}}\right) &\;,\; \mathbf{y} = \mathcal{S}\left(\mathbf{F}(\mathcal{X})\right)_{\alpha} \end{aligned} \tag{2}$$

where $\mathcal{X}, \hat{\mathcal{X}}$ are the input partial SRF and the ground truth whole SRFs respectively (as defined in Eq (1)), and α_{dense} is the density threshold distinguishing dense voxels from the air (usually set to 0). The sampling function $\mathcal{S}$ samples points in the grid space where the loss is defined on the outputs $\mathbf{y}$ and the ground truth densities' labels $\hat{\mathbf{y}}$. One of the challenges in working with sparse voxels of high resolution is that training the pipeline can not involve densifying the voxels to the original resolution due to prohibitive memory requirements. The input/output topologies are not necessarily the same, as the sparse convolutional strides and pruning can alter the sparse voxels' coordinates. This is why the sampling function $\mathcal{S}$ in Eq (2) is of utmost importance in guiding the training of SuRFNet. We sample the loss according to Quantized Gaussian sampling (Q-Gaussian) with random coordinates centered at the middle of the voxel grid. Simply put, the Q-Gaussian is 3D normal distribution quantized to integer coordinates to give a prior about where the output is expected and where the loss is defined. Further details and alternative configurations are provided in Section 6.1.

Radiance color loss. To ensure the output SRFs follow the ground truth optimized SRFs in radiance color, we follow the simple L1 loss on the radiance colors $\boldsymbol{\rho}$. However, as mentioned earlier, some of the non-empty voxels contain low density and will not be seen in the rendering, and can contain any random colors. Therefore, we mask these non-empty

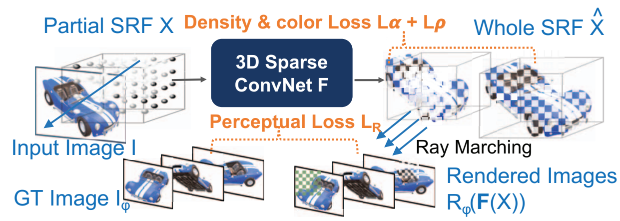

Figure 8. **SuRFNet: Learning to Generate Whole Radiance Fields from Partial Views.** We process the input images into partial SRFs $\mathcal{X}$ before learning a sparse convolutional network to generate the whole SRF. A perceptual loss is employed on the rendered images from poses ϕ to enhance the perceptual quality of the generated SRF. The whole SRF $\hat{\mathcal{X}}$ is used to 3D-supervise the SRF generation with density and radiance color losses.

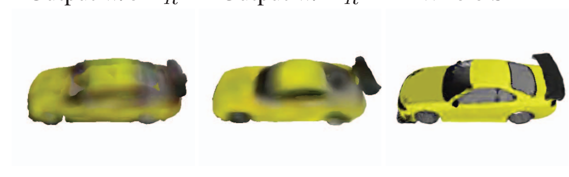

Figure 9. **Effect of the Perceptual Loss L_R.** Adding a perceptual loss on volume-rendered images during training SuRFNet insures the rendered images remain closer to how they should be rendered, as the 3D radiance colors supervision won't guarantee the rendering quality. *(left)*: without perceptual loss , *(middle)*: with the loss.

low-density voxels out of the L1 loss as follows:

$$\begin{aligned} L_\rho\left(\mathcal{X}\,,\,\hat{\mathcal{X}}\right) &= \|\mathbf{M}_\alpha \mathbf{F}(\mathcal{X})_\rho - \mathbf{M}_\alpha \hat{\mathcal{X}}_\rho\|_1 \\ \text{s. t.}\quad \mathbf{M}_\alpha &= \mathbb{1}(\hat{\mathcal{X}}_\alpha > \alpha_{\text{dense}}) \end{aligned} \tag{3}$$

Perceptual loss. Using only the 3D radiance color loss in Eq (3) ignores the rendering quality of the generated SRF, and would make it sensitive to hyperparameters (see Figure 9). Hence, we introduce an online perceptual loss that would volume render the generated SRF during training from M random views that come from the same ground truth image poses ϕ, and an L1 loss is defined between the generated images and the ground truth posed images $\mathbf{I}_\phi$.

$$L_R\left(\mathcal{X}\right) = \|\mathcal{R}_\phi\left(\mathbf{F}(\mathcal{X})\right) - \mathbf{I}_\phi\|_1, \tag{4}$$

where $\mathcal{R}_\phi$ is the fast volume rendering function that renders SRFs from poses ϕ using the trilinear interpolation between voxels proposed in Plenoxels[14].

The final loss to train the network $\mathbf{F}$ would be combining the three losses in Eq (2,3,4) as follows:

$$\text{Loss}_\mathbf{F} = L_\alpha + \lambda_\rho L_\rho + \lambda_R L_R, \tag{5}$$

where λ_ρ, λ_R are hyperparameters to control the radiance colors compared to the density predictions. The network is trained on all N whole SRFs $\hat{\mathcal{X}}$ in the dataset, while the input SRFs $\mathcal{X}$ are randomly chosen from several partial SRFs created by the same number of images from those shapes.

5. Experiments

5.1. Collecting SPARF

The engineering aspect of collecting, storing, and organizing the one million SRFs with multiple resolutions is as challenging as training properly on SRFs. In order to do that in manageable time and memory, while maintaining high quality in the optimized samples, a set of strategies is employed. The dimension of the radiance color d is chosen to be $d = 3 \times 4 = 12$ of 4 spherical harmonics factors of RGB channels for view-dependent SRF and $d = 3 \times 1 = 3$ for fixed RGB colors of the SRF. Since the input partial SRFs are noisy, we use $d = 3$ while the final output SRFs use $d = 12$ for high-quality image generation. Using fewer Spherical Harmonics components (from 9 to 4 per RGB channel) reduces the optimization space by 40% and time by 10%, while maintaining the same PSNR. Using RGB as colors instead of SH factors reduces PSNR by ~ 1 dB, space by 80%, and time by 20%. Running Plenoxels [14] for fewer iterations ($3\times$12K) reduces the time by 30% while maintaining the same PSNR. Using 400 views/shapes in SPARF to optimize the SRFs keep the time manageable in optimization (~ 4 minutes for the 512 resolution whole SRFs) while maintaining high PSNR (~ 30dB). It takes way less time than that for the rest of the setups (lower resolution or partial SRFs). Other similar sparse voxels optimizations (*e.g.* InstantNGP [43]) can be used to obtain SRFs similarly as fast with similar quality. A total of four variants of the partial SRFs are collected for all the resolutions and the partials use 1 and 3 images. The anatomy of the distribution of classes and SRFs in SPARF is presented in Figure 7. More details about SPARF and visualizations of some of its samples are available in the supplementary material.

5.2. Training Setup

Dataset. We pick our SPARF for the task of predicting whole SRFs for the purpose of novel view synthesis. The other datasets (SRN [56] and NMR [46]) are too small or low in resolution, which prevents optimizing high-quality radiance fields (see Figure 3).

Evaluation metrics. Following the previous novel view synthesis works [71, 14, 32], we use PSNR, SSIM, and LPIPS [74] as metrics to evaluate the synthesis. However, one key difference between our work and previous ones is that our setup is a learning setup (with training and validation), while previous works treat it as an optimization problem. Most previous works on novel view synthesis try to only generalize the generated views on the *same* shape, while we aim to generalize *across shapes of the same category* and *across views*. We treat the collected whole SRFs as ground truth labels for the input few images from the training set. We

Baselines	SPARF Validation PSNR chair	watercraft	rifle	display	lamp	speaker	cabinet	bench	car	airplane	sofa	table	phone	mean
Plenoxels [14] (1V)	9.2	11.1	11.7	8.0	13.6	8.2	10.4	10.5	7.1	12.8	9.3	9.9	8.3	10.0
Plenoxels [14] (3V)	10.7	13.3	14.9	9.7	15.8	10.4	12.4	11.6	7.1	14.6	11.6	10.8	9.7	11.7
PixelNerf [71] (1V)	13.3	16.3	16.7	11.9	17.6	11.3	14.5	14.6	13.2	19.2	13.3	13.2	13.2	14.5
PixelNerf [71] (3V)	13.5	16.6	16.9	12.2	17.9	11.9	14.9	14.8	13.4	**19.4**	13.4	13.3	13.3	14.7
VisionNeRF [32] (1V)	13.0	15.6	15.8	11.7	16.7	11.2	14.0	14.3	12.7	17.8	13.3	13.0	12.6	14.0
SuRFNet (ours) (1V)	11.6	16.2	17.0	12.0	16.2	12.6	17.0	13.5	16.6	17.5	14.1	10.1	15.3	14.6
SuRFNet (ours) (3V)	**15.3**	**18.3**	**18.8**	**15.0**	**19.0**	**16.6**	**20.0**	**15.6**	**16.6**	18.5	**18.1**	**14.9**	**17.8**	**17.3**

Table 2. **SPARF Benchmark on Out-Of-Distribution View Synthesis**. We compare the validation PSNR of some of the widely used novel view synthesis techniques on the SPARF dataset for the generalization of novel view synthesis beyond a single example and on view tracks completely different from the ones seen in training views. One view (1V) and three views (3V) inputs are reported.

Figure 10. **SuRFNet: Generating High-Resolution Radiance Fields**. We show some volume-rendered sequences based on our SuRFNet voxel radiance field outputs , given only 3 input images.

consider the validation PSNR, SSIM, and LPIPS of the input images at the validation SRF set of shapes (on the test images of those shapes) as the main evaluation metrics. Also, we report validation accuracy = $\frac{\text{validation PSNR with test few images}}{\text{whole SRF optimization's PSNR}}$ and propose it as a new metric to evaluate such a learning setup of SRFs. Furthermore, as we describe in Section 3, SPARF has 10 posed Out-Of-Distribution (OOD) images for every 3D shape to evaluate the robustness of novel view synthesis methods in the unconstrained setup. We report these *OOD* PSNR, SSIM, LPIPS, and Accuracies as well.

Basleines. We use PixelNeRF [71] (ResNet34 backbone), Plenoxels [14], and VisionNerf [32] (ViT-B backbone) as the main baselines for our work. Our SuRFNet network has two sizes: large (87 million parameters) and small (13 million parameters).

5.3. Results

We show qualitative results of generating novel views from few input images on unseen shapes in Figure 10. We also show qualitative comparisons in Figure 12. We present a summary of the quantitive evaluations next, where SuRFNet achieves state-of-the-art results on unconstrained novel view synthesis from one or few images on unseen shapes.

SPARF View-Generalization benchmark. In Table 2, we report the average PSNR results on the validation set of SPARF for different methods on unseen shapes during training on all 13 different object classes and on out-of-distribution views. It shows that our SuRFNet can generalize to out-of-distribution views on unseen shapes during test time, surpassing state-of-the-art PixelNeRF [71] and VisionNeRF [32]. Visual comparisons can be found in Figure 12. As can be seen from those results, the learned 3D prior results in multi-view consistency, especially when rendering from out-of-distribution views.

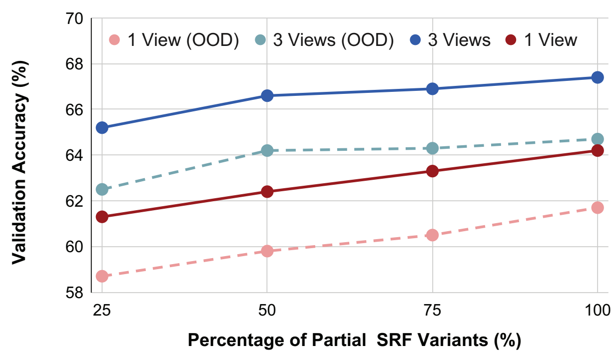

Figure 11. **Scaling-Up Training on SRFs**. As the training data (partial SRFs) of radiance fields increase, the generalization improves, as can be seen in the car class here. The 3-view and 1-view metrics are reported with test and OOD metrics.

6. Analysis and Insights

6.1. Ablation Study

Effect of Dataset Size. We study the effect of increasing the dataset size (Partial SRFs) on the generalization performance of SuRFNet in Figure 11. It shows that as the size increase (normalized by the total number of shapes in *car* class), the generalization performance increase. This scalability effect underlines the importance of SPARF. This justifies collecting 4 variants per resolution (as detailed in Figure 7).

Training SuRFNet. We ablate different components of SuRFNet's architectures and the loss configuration choices and report the results in Table 3. The Results show that increasing the size of the network (from 13M to 87M parameters) helps improve generalization accuracy. Also, they show the importance of the loss components proposed in Eq (5). The use of only density loss creates a reasonably dense shape but without colors. While combining the density loss with the 3D radiance color loss creates colorful objects, it does not perform well in the task of novel view synthesis as it does not respect how the object renders. Small radi-

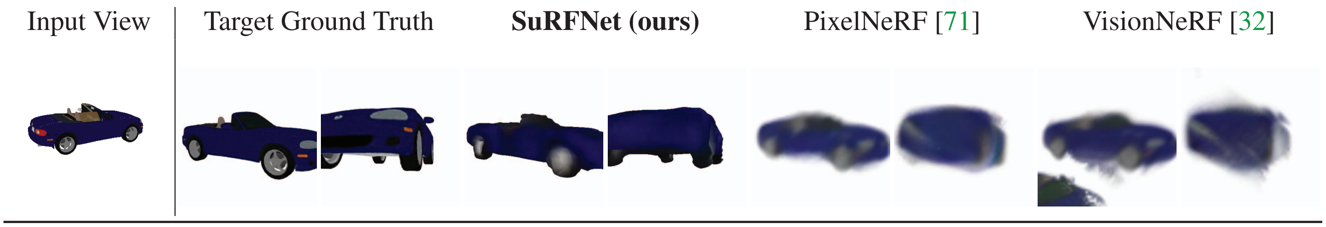

Figure 12. **Qualititve Comparisons**. We show different renderings from our SuRFNet outputs generated from a single image compared to other methods (pixel-NeRF [71], and VisionNeRF [32]) and whole SRF "GT" renderings. Note that the predicted two views lay outside the training views distribution (zoomed in randomly). This test highlights the weakness of the 2D-based baselines [71, 32] outside the training track, while our 3D approach maintains multi-view consistency everywhere.

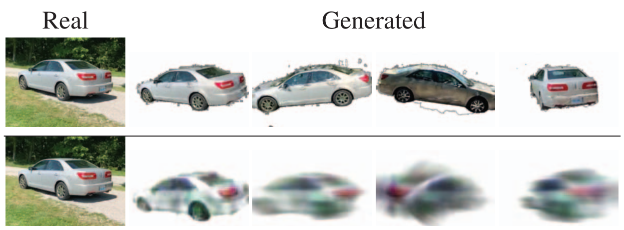

Figure 13. **Real images** We show real images of Co3D [52] and the corresponding generated views from our SRFs (*top row*) and pretrained PixelNeRF [71] (*bottom row*).

3D Backbone		**Loss components**			**Results**
Small	Large	(L_α)	(L_ρ)	(L_R)	Val. Acc.
✓	-	✓	✓	-	65.2
✓	-	✓	✓	✓	65.7
-	✓	✓	✓	-	66.4
-	✓	✓	✓	✓	**68.2**

Table 3. **Ablation Study**. We ablate different components of in SuRFNet (3D backbone and SRF-Loss) and report validation accuracy of *car* class.

ance distortions lead to large image errors (see Figure 9 for the importance of the perceptual loss). More ablations on the network, loss sampling, and hyperparameters of training SuRFNet can be found in supplementary material.

6.2. Testing on Real Images

Previous works on similar large-scale learning setups report only the ShapeNet benchmark [32, 53, 61] for standardized evaluations. Testing on real images can provide additional value. Hence, In Fig. 13, we show renderings of whole SRFs (used in our SuRFNet pipeline) trained on *real* Co3D [52] images *vs*. pretrained PixelNeRF renderings. This demonstrates the potential for the proposed method on real images. The small noise is due to background corruption resulting from imperfect masks provided in Co3D.

6.3. Speed and Compute Cost

To assess the contributions of the SuRFNet pipeline, we study the time and memory requirements of each element in the pipeline. We record in Table 4 the number of floating-

Network	Network FLOP (G)	Network Infer. (ms)	Parameters Number (M)	Rendering Speed (FPS)
PixelNeRF [71]	7.3	5.33	21.8	1.2
VisionNerf [32]	33.7	12.5	68.6	1.2
SuRFNet (small)	~15	14.4	13.4	15
SuRFNet (large)	~100	90.0	87.3	15

Table 4. **Time and Memory Requirements**. We assess the computational cost of the main components studied

point operations (GFLOPs) and the runtime of a forward pass (including rendering) for a single output image from one input image on Titan RTX GPU. Speed and compute details about collecting the SRFs are included in Section 5.1.

7. Conclusions and Future Work

We propose a large-scale dataset SPARF of sparse radiance fields that include around one million SRFs and 17 million posed images of 3D shapes. The dataset aims to move the community in the direction of treating radiance fields as a 3D data structure, instead of optimization results and MLP fitting. We domnstrated SPARF's utility using our proposed SuRFNet pipeline.

One crucial limitation in this work is the large amount of compute and memory necessary to create, store, and process SRFs, especially at high-resolution voxel grids. This creates a bottleneck in training, developing, and building on SRFs. Developing efficient methods to work and learn from sparse voxel grids would be a viable plan moving forward in order to develop deep and large models in this space.

Acknowledgement. This work was supported by the ERC Starting Grant Scan2CAD (804724) and the German Research Foundation (DFG) Research Unit "Learning and Simulation in Visual Computing". It was also supported by the KAUST Office of Sponsored Research through the Visual Computing Center funding. Part of the support is also coming from the KAUST Ibn Rushd Postdoc Fellowship program.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. *International Conference on Machine Learning (ICML)*, 2018. 1, 2

[2] Alex Holkner et al. Pyglet. 3

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

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2, 3

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 3

[7] Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Advances in Neural Information Processing Systems*, pages 9609–9619, 2019. 1

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 4, 5

[9] Angela Dai and Matthias Nießner. Scan2mesh: From unstructured range scans to 3d meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5574–5583, 2019. 2

[10] Dawson-Haggerty et al. trimesh. 3

[11] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13142–13153, 2023. 3

[12] 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. *ICLR*, 2021. 3

[13] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. 3

[14] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, June 2022. 1, 2, 3, 4, 5, 6, 7

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

[16] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9785–9795, 2019. 1, 2

[17] Shubham Goel, Georgia Gkioxari, and Jitendra Malik. Differentiable stereopsis: Meshes from multiple views using differentiable rendering. *arXiv preprint arXiv:2110.05472*, 2021. 1, 2

[18] Pengsheng Guo, Miguel Angel Bautista, Alex Colburn, Liang Yang, Daniel Ulbricht, Joshua M Susskind, and Qi Shan. Fast and explicit neural view synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3791–3800, 2022. 3, 5

[19] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, October 2021. 1

[20] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Voint cloud: Multi-view point cloud representation for 3d understanding. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[21] Rana Hanocka, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *arXiv preprint arXiv:2005.11084*, 2020. 2

[22] Qixing Huang, Xiangru Huang, Bo Sun, Zaiwei Zhang, Junfeng Jiang, and Chandrajit Bajaj. Arapreg: An as-rigid-as possible regularization loss for learning deformable shape generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5815–5825, 2021. 1

[23] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dex-nerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*, 2021. 3

[24] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 2, 3

[25] Yoonwoo Jeong, Seungjoo Shin, Junha Lee, Chris Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. PeRFception: Perception using radiance fields. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[26] Danilo Jimenez Rezende, S. M. Ali Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. In D. D. Lee,

M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4996–5004. Curran Associates, Inc., 2016. 1

[27] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3

[28] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems (NIPS)*, pages 2539–2547, 2015. 1

[29] Ruihui Li, Xianzhi Li, Ka-Hei Hui, and Chi-Wing Fu. Sp-gan: Sphere-guided 3d shape generation and manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021. 1, 2

[30] Zuoyue Li, Tianxing Fan, Zhenqiang Li, Zhaopeng Cui, Yoichi Sato, Marc Pollefeys, and Martin R Oswald. Compnvs: Novel view synthesis with scene completion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 447–463. Springer, 2022. 3

[31] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 4

[32] Kai-En Lin, Lin Yen-Chen, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 2, 3, 5, 6, 7, 8

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 1

[34] 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. 1, 2

[35] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, aug 1987. 4

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 1, 3

[37] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360{\deg} reconstruction of any object from a single image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[39] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. *arXiv preprint arXiv:2112.03221*, 2021. 1, 2

[40] 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)*, 38(4):1–14, 2019. 3

[41] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 4

[42] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3971–3980, June 2022. 1, 3

[43] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 1, 2, 3, 6

[44] Charlie Nash, Yaroslav Ganin, SM Ali Eslami, and Peter Battaglia. Polygen: An autoregressive generative model of 3d meshes. In *International Conference on Machine Learning*, pages 7220–7229. PMLR, 2020. 1, 2

[45] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2, 3

[46] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2, 3, 6

[47] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[48] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1, 3

[49] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[50] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 704–720, 2018. 1, 2

[51] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[52] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10901–10911, October 2021. 2, 3, 8

[53] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021. 2, 3, 8

[54] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[55] Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1

[56] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3, 6

[57] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, et al. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *arXiv preprint arXiv:2205.07058*, 2022. 2, 3

[58] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1, 2

[59] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3

[60] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3

[61] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. 8

[62] Xingkui Wei, Zhengqing Chen, Yanwei Fu, Zhaopeng Cui, and Yinda Zhang. Deep hybrid self-prior for full 3d mesh generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5805–5814, 2021. 2

[63] Mason Woo, Jackie Neider, Tom Davis, and Dave Shreiner. *OpenGL Programming Guide: The Official Guide to Learning OpenGL, Release 1*. Addison-wesley, 1998. 3

[64] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 2

[65] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020. 3

[66] Jun Yang, Yizhou Gao, Dong Li, and Steven L Waslander. Robi: A multi-view dataset for reflective objects in robotic bin-picking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9788–9795. IEEE, 2021. 3

[67] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 1, 2

[68] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 3

[69] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2

[70] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 1, 3

[71] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 1, 2, 3, 5, 6, 7, 8

[72] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. *arXiv preprint arXiv:1811.12328*, 2018. 1

[73] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[75] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3