

NeSF: Neural Semantic Fields for Generalizable Semantic Segmentation of 3D Scenes

Suhani Vora^{1*} Noha Radwan^{1*} Klaus Greff¹ Henning Meyer¹ Kyle Genova¹
Mehdi S. M. Sajjadi¹ Etienne Pot¹ Andrea Tagliasacchi^{1,2} Daniel Duckworth¹

¹Google Research ²University of Toronto

Abstract

We present NeSF, a method for producing 3D semantic fields from posed RGB images alone. In place of classical 3D representations, our method builds on recent work in implicit neural scene representations wherein 3D structure is captured by point-wise functions. We leverage this methodology to recover 3D density fields upon which we then train a 3D semantic segmentation model supervised by posed 2D semantic maps. Despite being trained on 2D signals alone, our method is able to generate 3D-consistent semantic maps from novel camera poses and can be queried at arbitrary 3D points. Notably, NeSF is compatible with any method producing a density field, and its accuracy improves as the quality of the density field improves. Our empirical analysis demonstrates comparable quality to competitive 2D and 3D semantic segmentation baselines on complex, realistically-rendered synthetic scenes. Our method is the first to offer truly dense 3D scene segmentations requiring only 2D supervision for training, and does not require any semantic input for inference on novel scenes. We encourage the readers to visit the [project website](#).

1. Introduction

High-level semantic understanding of 3D scenes as captured by digital images and videos is a fundamental objective of computer vision. Well-studied tasks such as scene classification [72], object detection [81], semantic segmentation [75], and instance segmentation [49] infer semantic descriptions of scenes from RGB and other sensors and form the foundation for applications such as visual navigation [7] and robotic interaction [6].

The most common approach to scene understanding is to narrow the scope to 2D (image-space) reasoning, wherein classical image-to-image architectures [118] are trained on

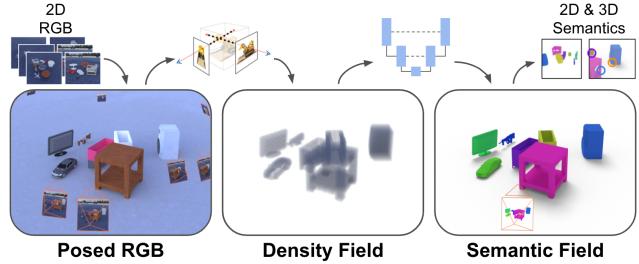


Figure 1. Overview – We train our method on collections of posed 2D RGB images and 2D semantic maps, each collection describing an independent scene. Given a new set of posed 2D RGB images, we extract an implicit volumetric representation of the scene’s 3D geometry and infer a 3D semantic field. The semantic field can then be used to render dense 2D semantic maps from novel camera poses or queried directly in 3D. Our method generalizes to novel scenes, and requires as little as one semantic map per scene at training time. We encourage readers to visit the [project website](#).

large collections of semantically-annotated images [78]. These methods, while straightforward to apply, only produce per-pixel annotations and largely ignore the underlying 3D structure of the scene. Instead, our goal is to use a set of RGB images with known poses to produce a 3D *semantic field*: a function mapping 3D positions to probability distributions over semantic categories.

For 3D semantic segmentation, most prior work relies on 3D sensors [36, 59, 71] and/or 3D semantic labels [9, 30]. Convolutional network architectures have been designed for 3D point clouds [23, 112], voxel grids [131], and polygonal meshes [51, 87]. However, 3D sensors are not as affordable nor as widely available as their RGB cameras, and 3D annotations are significantly more challenging to produce than their 2D counterparts and are hence generally scarce [41].

To overcome this challenge, researchers have adopted *hybrid* 2D/3D reasoning, so as to propagate densely supervised semantic signals from 2D projections back to an underlying 3D substrate [31, 41, 73]. At test time, these methods still

*Denotes equal contribution.

require a classical 3D representation to be provided as input, hence limiting their applicability and performance. One interesting exception is Atlas [97], which only requires posed photos at test time but still requires 3D supervision to train.

In parallel to these developments, a new family of 3D representations have emerged based on implicit, coordinate-wise functions [19, 91, 105]. In this regime, one trains a neural network to predict quantities such as occupancy, signed distance, density and radiance. Unlike explicit representations of geometry, neural implicit methods are memory efficient and are able to capture impressive levels of geometric detail [128]. However, the vast majority of methods built on this approach target computer graphics applications such as novel view synthesis [80, 114, 136, 142, 146] without an interpretable semantic understanding of the scene. One notable exception is Semantic-NeRF [149], which regresses a per-3D point semantic class in addition to radiance and density. Similar to NeRF [94], this method is only applicable to novel *views* within the same scene and does not provide the form of generalization one expects in classical semantic segmentation: the ability to infer semantics on novel *scenes*.

We introduce *Neural Semantic Fields (NeSF)*, a method for semantic segmentation of 3D scenes via image-space semantic annotations; see Figure 1. Unlike Semantic-NeRF, NeSF generalizes to scenes unobserved at training time. In place of an explicit scene representation, NeSF builds on the implicit representations of geometry recovered by methods such as NeRF. In particular, we apply a neural network to NeRF’s density field to obtain what we refer to as a scene’s *semantic field*. A semantic field is thus defined as a coordinate-wise function mapping 3D points to probability distributions over semantic categories. Similar to Semantic-NeRF, we apply the volumetric rendering equation to generate 2D semantic maps, enabling supervision from posed semantic annotations in image-space. To the best of our knowledge, NeSF is the first method capable of producing dense 2D and 3D segmentations of novel scenes from posed RGB images alone. The ability to reason about 3D information from 2D supervision alone is essential to the deployment of 3D computer vision at scale; while 2D sensors are ubiquitous, 3D sensors are expensive, unwieldy, and unlikely to be deployed in the mass market.

As large scale datasets of 3D semantically annotated scenes are scarce, we propose three novel datasets of increasing complexity: KLEVR, ToyBox5, and ToyBox13. While datasets for 2D and 3D semantic scene understanding already exist [10, 12, 30, 125], they lack the scale, detail, realism, and precision necessary to simultaneously evaluate 2D and 3D semantic segmentation. We construct over 1,000 scenes of randomly-placed, toy-sized objects and render hundreds of RGB images with realistic lighting and materials of each. Notably, random object placement breaks relational consistencies between objects that exist in available datasets,

enabling harder tasks. Each RGB image is paired with corresponding ground-truth camera intrinsics and extrinsics, a semantic map and a depth map. We evaluate our method on these three datasets and compare its performance to competitive techniques in 2D and 3D scene understanding.

Contributions.

- We introduce the first method for generating 3D semantic fields for novel scenes trained solely on posed RGB images and semantic maps. Unlike prior work, our method (i) can be queried anywhere within a bounded 3D volume, (ii) is capable of rendering semantic maps from novel camera poses, and (iii) generalizes to novel scenes with as few as one semantic map per scene at training time.
- We propose three new synthetic datasets for 2D and 3D semantic scene understanding. In total, these datasets comprise of over 1,000 scenes and 3,000,000 realistically-rendered and semantically annotated frames. Upon publication, these datasets will be released to the community alongside code to reproduce them.

2. Related works

We now briefly overview related work in semantic segmentation [52, 96] and 3D reconstruction [64].

Semantic segmentation. Semantic segmentation is a heavily researched area, with most methods targeting a fully supervised, single modality problem (2D [4, 16, 17, 22, 83, 119] or 3D [14, 21, 29, 47, 82, 95, 147, 150]). 2D approaches like DeepLab [16] train a CNN to segment each pixel in an image. There are also analogous approaches in 3D for various shape representations – point clouds [56, 110, 112, 113, 122, 129, 135], sparse or dense voxel grids [24, 25, 31, 43, 50, 117, 123], or meshes [51, 57]. In contrast to these methods, our method reconstructs and then segments a dense 3D representation from 2D inputs and supervision alone, and does not require ground truth 3D annotations or input geometry.

Hybrid and multi-modal methods. Many methods use one data modality to supervise or inform another [1, 3, 37, 38, 42, 48, 65, 68, 69, 76, 93, 98, 104, 130, 148]. For 3D semantic segmentation, multiview fusion [2, 55, 73, 84, 86, 89, 133, 133, 145] is a popular family of methods that require only image supervision. However, these methods reason exclusively in the image domain and require an input 3D substrate such as a point cloud or polygonal mesh on which to aggregate 2D information. Similarly, Genova *et al.* [41] propose a method for 3D point cloud segmentation from 2D supervision, but still requires input 3D geometry. In a separate line of work, researchers have proposed pipelines for 3D segmentation that benefit from 2D image features [31, 62, 73, 74, 134]. Unlike our method, these approaches also require a full 3D supervision.

Implicit representations. Most similar to our approach, Atlas [97] learns a 3D implicit TSDF reconstruction from

2D images while also learning to segment the predicted scene geometry. However, this approach requires ground truth 3D data and supervision, while our method requires only images at *both* train and test time. Other methods use implicit representations to reconstruct a 3D scene [109, 126] or shape [11, 18, 20, 33, 34, 40, 92, 101, 103, 106, 108, 111, 120, 128]. A more recent work approaches the problem with image supervision only [126] but does not consider semantics.

Neural radiance fields. Recently, a variety of methods based on NeRF [94] have become popular for novel view synthesis [5, 39, 53, 63, 77, 79, 80, 85, 99, 107, 115, 124, 136, 141], 3D reconstruction [8, 8, 15, 27, 35, 54, 60, 61, 102, 116, 132, 139, 142, 143], generative modeling [70, 90, 100, 121] and semantic segmentation [149]. The majority of these models demonstrate impressive results on novel view synthesis but are only applicable in the single-scene setting. Others generalize to novel scenes but focus on novel view synthesis or reconstruction. In contrast, ours is the first approach capable of generating 3D semantic segmentations of novel scenes without supervision at test time.

3. Method

We train NeSF on a collection of S scenes, each described by a collection of RGB images $\{\mathcal{C}_{s,c}^{\text{gt}} \in [0, 1]^{H \times W \times 3}\}$ and paired to a collection of semantic maps $\{\mathcal{S}_{s,c}^{\text{gt}} \in \mathbb{Z}_+^{H \times W}\}$. Both images and maps are indexed by camera index c and a scene index s . For the sake of exposition, we assume that each RGB map is paired with a semantic map and that each scene contains C map pairs, but the method itself makes no such assumption. Similar to prior work, we also assume the availability of camera calibration parameters $\{\gamma_{s,c} \in \mathbb{R}^\Gamma\}$ providing an explicit connection between each pixel and the 3D ray \mathbf{r} cast within the 3D scene. We consider the problem of jointly estimating the camera calibration and the scene representation outside the context of this work, see [77, 137, 140]. Our method involves two stages, which are described in the following subsections:

- **Section 3.1:** In the first stage, we pre-train neural radiance fields on posed RGB maps $\{(\mathcal{C}_{s,c}^{\text{gt}}, \gamma_{s,c})\}$ *independently* for each scene $s \in [1 \cdots S]$. This results in a set of neural radiance fields with network parameters $\{\theta_s\}$. To focus on the core task of understanding from 3D geometry, we disregard the radiance portion of these fields and employ the *volumetric density fields* $\sigma(\mathbf{x} | \theta_s) \in [0, \infty)$ below.
- **Section 3.2:** In the second stage, we train a density-to-semantics *translation* network \mathcal{T} parameterized by $\tau = \{\tau_{\text{unet}}, \tau_{\text{mlp}}\}$. Given a scene’s 3D geometry represented by density field $\sigma_s = \sigma(\cdot | \theta_s)$, this network produces a 3D semantic field $\mathbf{s}(\mathbf{x} | \sigma_s, \tau)$ assigning each point a probability distribution over semantic categories. While the translation network produces a 3D field, we apply the volumetric rendering equation to obtain 2D seman-

tic maps from reference camera poses $\{\gamma_{s,c}\}$. Predicted semantic maps can then be compared to their ground truth counterparts $\{\mathcal{S}_{s,c}^{\text{gt}}\}$ in a differentiable way.

3.1. NeRF pre-training

To extract an accurate, dense representation of each scene’s 3D geometry, we leverage neural radiance fields as proposed in [94]. To simplify notation, we drop scene index s for the remainder of this section as all scenes can be trained independently and in parallel. More specifically, given a collection of posed RGB images $\{(\mathcal{C}_c^{\text{gt}}, \gamma_c)\}$, and denoting with $\mathbf{r} \sim \mathcal{R}(\gamma_c)$ rays corresponding to pixels from image $\mathcal{C}_c^{\text{gt}}$, a neural radiance field model with parameters θ is trained by minimizing the squared photometric reconstruction loss:

$$\mathcal{L}_{\text{rgb}}(\theta) = \sum_c \mathbb{E}_{\mathbf{r} \sim \mathcal{R}(\gamma_c)} [\|\mathcal{C}(\mathbf{r} | \theta) - \mathcal{C}_c^{\text{gt}}(\mathbf{r})\|_2^2] \quad (1)$$

where $\mathcal{C}_c^{\text{gt}}(\mathbf{r})$ is the ground truth color of ray passing through a pixel in image c , and the color $\mathcal{C}(\mathbf{r} | \theta)$ is computed by applying the volumetric rendering equation with the ray’s near and far bounds $t \in [t_n, t_f]$:

$$\mathcal{C}(\mathbf{r} | \theta) = \int_{t_n}^{t_f} w(t | \theta) \cdot \mathbf{c}(t | \theta) dt \quad (2)$$

Let $\mathbf{r}(t) = \mathbf{o} + \mathbf{d}t$ represent a point along a ray with origin \mathbf{o} and direction \mathbf{d} . Weight $w(t) = w(\mathbf{r}(t))$ is then defined as:

$$w(t) = \underbrace{\exp \left(- \int_{t_n}^t \sigma(u) du \right)}_{\text{visibility of } \mathbf{r}(t) \text{ from } \mathbf{o}} \cdot \underbrace{\sigma(t)}_{\text{density at } \mathbf{r}(t)} \quad (3)$$

where the volumetric density $\sigma(t)$ and radiance fields $\mathbf{c}(t)$ are predicted by a multi layer perceptron (i.e. MLP) with Fourier feature encoding. We refer the reader to the original work [94] for further details and the discretization of these integrals [88].

Training. While neural radiance fields are acknowledged to be slow to train, we find that we are able to fit a single model to sufficient quality in ≈ 20 minutes on eight TPUv3 cores on the Google Cloud Platform. Once trained, the per scene parameters $\{\theta_s\}$ are held fixed.

3.2. Semantic Reasoning

We now present a method for mapping 3D density fields to 3D semantic fields. To recap, we train a *translation* model $\mathcal{T}(\sigma | \tau)$ to produce a semantic field $\mathbf{s}(\mathbf{x} | \sigma, \tau)$ being given access to the density field of a scene σ , where \mathbf{s} assigns a probability distribution over semantic categories at each 3D point in space. We optimize translation model’s parameters τ with 2D annotations alone. Our inspiration is drawn from methods that translate *explicit* representations of geometry into semantics [32, 112], and in observing that for

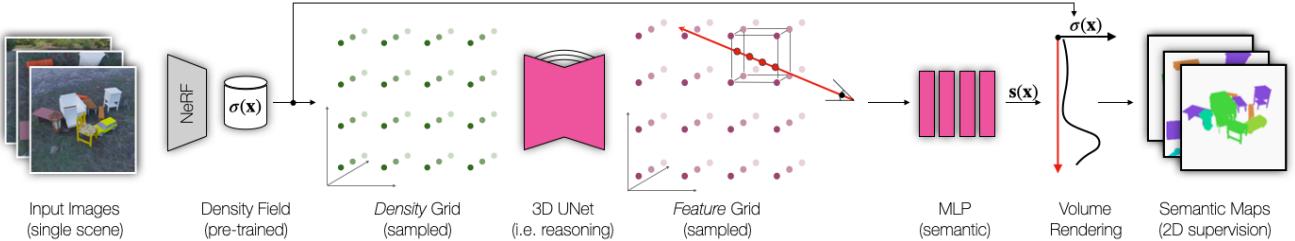


Figure 2. Architecture – Given a pre-trained NeRF model, we sample its volumetric density grid to obtain the 3D scene representation. This grid is converted to a semantic-feature grid by employing a fully convolutional volume-to-volume network thus allowing for geometric reasoning. The semantic-feature grid is in turn translated to semantic probability distributions using the volumetric rendering equation. Note the semantic 3D UNet is trained across all scenes in the TRAIN SCENES set, though not explicitly depicted for the sake of simplicity. Additionally, note that NeSF is trained *solely* using 2D supervisory signals and that no segmentation maps are provided at test time.

density fields provide an *implicit* notion of geometry. For ease of exposition, the overall architecture illustrate in Figure 2, is broken down into several discrete steps: ① density grid extraction, ② spatial reasoning, ③ feature decoding, ④ supervision, and ⑤ data augmentation.

Density grid extraction. Our method starts by uniformly evaluating the density field on a 3D lattice with spacing ϵ between samples:

$$\Sigma_s = \Sigma(\sigma_s) = \{\sigma(\mathbf{x} | \theta_s) \text{ s.t. } \mathbf{x} \in [-1 : \epsilon : +1]^3\} \quad (4)$$

While this operation limits the spatial resolution of the original density field, it presents a natural representation for further processing.

Spatial reasoning (3D). We apply a 3D UNet [26] to Σ_s to obtain a *feature* grid \mathcal{F}_s of the same spatial resolution as Σ_s :

$$\mathcal{F}_s = \mathcal{F}(\Sigma_s | \tau_{\text{unet}}) = \text{UNet3D}(\Sigma_s | \tau_{\text{unet}}) \quad (5)$$

This step is essential as a point-wise measurement of the density field $\sigma(\mathbf{x})$ does not contain sufficient information to capture 3D structure. After all, $\sigma(\mathbf{x})$ only measures the volumetric density at a point, while the 3D structure requires reasoning over local spatial neighborhoods. Note that we *share* translation network parameters τ_{unet} across scenes, enabling generalization to novel scenes not available at training time.

Feature decoding. Given a query point $\mathbf{x} \in \mathbb{R}^3$, we interpolate within the feature grid \mathcal{F}_s to obtain a feature vector corresponding to \mathbf{x} . We then employ a neural network decoder \mathcal{D} to generate a field of probability distributions over semantic categories:

$$\mathbf{s}(\mathbf{x} | \mathcal{F}_s, \tau_{\text{mlp}}) = \mathcal{D}(\text{TriLerp}(\mathbf{x}, \mathcal{F}_s) | \tau_{\text{mlp}}) \quad (6)$$

where \mathcal{D} is a multilayer perceptron with trainable parameters τ_{mlp} and TriLerp applies trilinear interpolation similar to [80]. Like their UNet counterpart, parameters τ_{mlp} are *shared* across all scenes.

Supervision. To supervise the training of parameters τ , we employ volumetric rendering as in NeRF [94] but adapt it to render semantic maps as in Semantic-NeRF [149]:

$$\mathcal{S}(\mathbf{r} | \sigma_s, \tau) = \int_{t_n}^{t_f} w(t | \sigma_s) \cdot \mathbf{s}(\mathbf{x} | \sigma_s, \tau) dt \quad (7)$$

We supervise the training process for τ by minimizing the softmax cross-entropy between rendered semantic and ground truth semantic maps along with a smoothness regularization term:

$$\mathcal{L}_{\text{sem}}(\tau) = \mathbb{E}_s [\mathcal{L}_{\text{sem}}(\theta_s, \tau)] \quad \text{where:} \quad (8)$$

$$\mathcal{L}_{\text{sem}}(\theta_s, \tau) = \sum_c \mathbb{E}_{\mathbf{r} \sim \mathcal{R}(\gamma_c)} [\text{CE}(\mathcal{S}(\mathbf{r} | \theta_s, \tau) - \mathcal{S}_c^{\text{gt}}(\mathbf{r}))]$$

We include an additional smoothness regularization term to encourage similar predictions in local neighborhoods. We sample points $\mathbf{x} \sim \text{Uniform}([-1, 1]^3)$ and normally distributed noise $\epsilon \sim \mathcal{N}(0, 0.01)$,

$$\mathcal{L}_{\text{reg}}(\tau) = \mathbb{E}_{\mathbf{x}, \epsilon} [| | \mathbf{s}(\mathbf{x} | \mathcal{F}, \tau) - \mathbf{s}(\mathbf{x} + \epsilon | \mathcal{F}, \tau) | |_2^2] \quad (9)$$

Our total loss is thus $\mathcal{L}(\tau) = \mathcal{L}_{\text{sem}}(\tau) + \lambda \mathcal{L}_{\text{reg}}(\tau)$.

Data augmentation. To increase the robustness of our method and similarly to classical methods [112, 113], we apply data augmentation in the form of random rotations around the z-axis (i.e. upwards). In particular, we randomly sample an angle $\gamma \in [0, 2\pi]$ at each step of training. Rather than extracting a density grid in NeRF’s original coordinate system, we construct a rotation transformation R_γ and query NeRF at points $\mathbf{x}' = R_\gamma(\mathbf{x})$, resulting in the following density grid:

$$\tilde{\Sigma}_s = \{\sigma(\mathbf{x}' | \theta_s) \text{ s.t. } \mathbf{x} \in R_\gamma^{-1}([-1 : \epsilon : +1]^3)\} \quad (10)$$

Note that this procedure does not necessitate the retraining of NeRF models.

	KLEVR	ToyBox5	ToyBox13
# scenes	100 / 20	500 / 25	500 / 25
# cameras/scene	210 / 90	210 / 90	210 / 90
# total cameras	36,000	1,575,00	1,575,00
frame resolution	256×256	256×256	256×256
# objects/scene	4-12	4-12	4-12
# object instances	5	25,905	39,695
# background instances	1	383	383

Table 1. **Dataset statistics** – Each dataset consists of a set of train and novel scenes, wherein each scene’s cameras are split into a train and test set (denoted by a “/”).

4. Datasets – Table 1 and Figure 3

To investigate NeSF, we require datasets describing the appearance and semantics of a large number of scenes from multiple points of view. While existing datasets based on indoor and self-driving sensor captures exist [10, 12, 30, 127], we desire a controlled setting where distractors such as motion blur, camera calibration error, and object motion can be eliminated. To this end, we introduce three new datasets built on Kubric [45]: KLEVR, ToyBox5, and ToyBox13. Each dataset consists of hundreds of synthetic scenes, each containing randomly-placed 3D objects which are photorealistically rendered by a path tracer [28] supporting soft shadows, physically based materials, and global illumination effects. Each scene is described by a set of posed frames, where each frame provides an RGB image, a semantic map, and a depth map rendered from a shared camera pose. We provide the Kubric worker script to generate such scenes, so as to enable follow-up research to build ever-more-challenging datasets.

KLEVR. We design the KLEVR dataset to be a simple testbed akin to MNIST in machine learning. Inspired by CLEVR [66], each scene contains 4 to 12 simple geometric objects placed randomly on a matte grey floor. Each object is assigned a random hue and scale, and is constrained to lie within a fixed bounding box. The semantic category of an object is set equal to its geometry class (e.g. cube, cylinder, etc). While only the shape of each object is semantically relevant, color, scale, and placement serve as distractors. For each scene, we render 300 frames from randomly-sampled camera poses aimed at the scene’s origin. Camera poses are constrained to lie in the upper hemisphere surrounding the scene. For each frame, we render an RGB image, a semantic map, and a depth map.

ToyBox5 and ToyBox13. These datasets are designed to imitate scenes of children’s bedrooms and are designed to be more challenging. Scenes are constructed from a large vocabulary of ShapeNet [13] objects coupled with HDRI backdrops (floor, horizon, and environment illumination) captured in the wild [144]. Like KLEVR, each scene consists of 4-12 randomly-placed objects, and frames are rendered

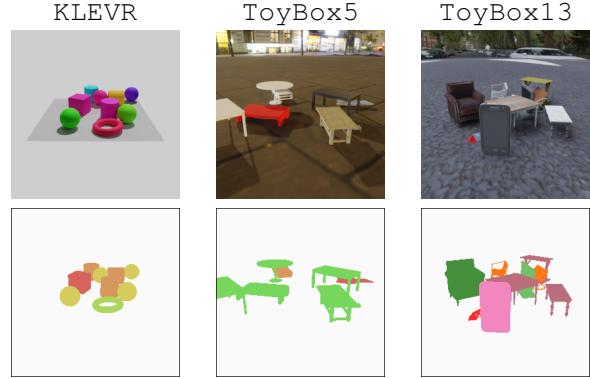


Figure 3. **Dataset examples** – Each frame includes an RGB image, semantic map, and depth map (not pictured here).

from 300 independently-sampled camera poses. Objects are sampled at random from the 5 and 13 most common object categories, respectively for ToyBox5 and ToyBox13. Such splits have been commonly used in the 3D deep learning literature [33, 40, 46, 92]. Like the objects themselves, backdrops are sampled at random when constructing a scene. With thousands of objects per category to choose from, most object instances appear rarely or only once.

Train/Test splits. To enable evaluation from novel views within the same scene, we randomly partition each scene’s frames into TRAIN CAMERAS and TEST CAMERAS; the latter representing the set typically used to evaluate methods in novel view synthesis [94]. For evaluation across scenes, we further partition scenes into TRAIN SCENES and NOVEL SCENES.

5. Experiments

We evaluate NeSF on the three datasets described in Section 4. Unless otherwise stated, we train NeRF models on all TRAIN CAMERAS from all TRAIN SCENES. To mimic a label-scarce regime, we choose to provide NeSF supervision from semantic maps corresponding to 9 randomly-chosen cameras per scene. For 2D evaluation, we randomly select 4 cameras from each NOVEL SCENES’ TRAIN CAMERAS. For 3D evaluation, we use camera parameters and ground truth depth maps to derive a labeled 3D point cloud from the same 4 cameras. Semantic segmentations are evaluated according to 2D and 3D mean intersection-over-union. Further details are provided in the [supplementary material](#).

Training Details. Each scene is preprocessed by training an independent NeRF for 25k steps with Adam using an initial learning rate of $1e-3$ decaying to $5.4e-4$ according to a cosine rule. Our NeRF architecture follows the original work. NeSF is trained for 5k steps using Adam with an initial learning rate of $1e-3$ decaying to $4e-4$. As input for NeSF, we discretize density fields by densely probing with $\epsilon=1/32$

	TRAIN CAMERAS		TEST CAMERAS	
	2D mIoU	3D mIoU	2D mIoU	3D mIoU
NeSF	92.7	97.8	92.6	97.5
DeepLab [16]	97.1	N/A	N/A	N/A
SparseConvNet [44]	N/A	99.7	N/A	99.7

Table 2. **Quantitative comparison on KLEVR** – NeSF is competitive with 2D and 3D baselines. See additional details in Table 3.

resulting in 64^3 evenly-spaced points in $[-1, +1]^3$. This density grid is then processed by the 3D UNet architecture of Çiçek *et al.* [26] with 32, 64, and 128 channels at each stage of downsampling. The semantic latent vector is processed by a multilayer perceptron consisting of 2 hidden layers of 128 units. Our models are trained on 32 TPUv3 cores.

Segmentation baselines (2D/3D). We compare NeSF to two popular semantic segmentation baselines, DeepLab [16] and SparseConvNet [44]. DeepLab follows a traditional 2D semantic segmentation pipeline, producing segmentation maps from RGB images. We train DeepLab v3 with Wide ResNet [138] for $55k$ steps on 16 TPUv3 chips. SparseConvNet is a point cloud segmentation method and, unlike NeSF and DeepLab, requires explicit 3D supervision. We train SparseConvNet asynchronously on 20 NVIDIA V100 GPUs with momentum using a base learning rate of $1.5e-2$ and decaying to 0 over the final $250k$ steps of training. We refer the reader to Section 5.1 and the [supplementary material](#) for further details.

5.1. Comparisons to baselines

Our first set of experiments evaluates the performance of our proposed method in comparison to alternative benchmark methods on the KLEVR, ToyBox-5, and ToyBox-13 datasets. As far as we are aware, NeSF is the *first* method capable of simultaneously producing 3D geometry, 2D semantic maps, and 3D semantic labels directly from posed RGB images at inference time. Unlike prior work [97], our method is trained on posed 2D supervision alone. As no existing method is directly comparable, we compare NeSF to competitive baselines for 2D image segmentation and 3D point cloud segmentation.

Comparison to DeepLab [16] (2D). To maintain a fair comparison in 2D, we train both the semantic phase of NeSF and DeepLab on an identical set of paired RGB images and semantic maps for a fixed set of scenes (i.e. 9 per scene in TRAIN SCENES). NeSF has further access to all 210 RGB maps associated with each scene’s TRAIN CAMERAS, which are used to fit per-scene NeRF models. Both methods are evaluated on a random sampling of frames from NOVEL SCENES, 4 per scene. To emphasize the 3D nature of NeSF, we evaluate on additional camera poses from NOVEL SCENES where RGB information is not available, an additional 4 per scene from each scene’s TEST CAMERAS.

Comparison to SparseConvNet [44] (3D). As SparseConvNet requires 3D input, we derive an oracle point cloud for each scene from camera poses and *ground truth* depth maps – hence giving an *unfair advantage* to this method, establishing an upper bound on performance given full 3D supervision. For this, we use the same 210 train frames used to fit NeRF models in the 2D comparison. We further select a subset of each point cloud for 3D semantic supervision; namely, the points corresponding to the 9 semantic maps supervising NeSF and DeepLab. We evaluate NeSF and SparseConvNet on two sets of 3D points on NOVEL SCENES. The first set is a subset of each point cloud corresponding to 4 randomly chosen frames from each scene’s TRAIN CAMERAS. These points are available to SparseConvNet as part of each scene’s 3D representation. The second set is a set of additional query points derived from 4 additional frames from each scene’s TEST CAMERAS. As SparseConvNet is not designed to classify points beyond its input point cloud, we apply a nearest neighbor label propagation procedure to assign labels to the latter.

Quantitative comparisons – Table 2 and Table 3. While all methods perform comparably on the KLEVR dataset, model quality varies drastically on the more challenging datasets. On ToyBox5, our method performs comparably to DeepLab, but on ToyBox13, it underperforms by 6.6% in 2D mIoU. While our method does not achieve the same level of accuracy as DeepLab on frames where RGB images are available, it is able to achieve near identical accuracy on *novel* camera poses, a task DeepLab is unable to approach. In order to focus on the foundational properties of our method, we have chosen to limit NeSF to 3D geometric information alone. Incorporation of 2D information via projection onto ground truth cameras in the spirit of PixelNeRF [142] or IBRNet [136] is straightforward.

As expected, our method also underperforms SparseConvNet by 4.7-5.2% on ToyBox5 and 19.8-23.1% on ToyBox13. Unlike SparseConvNet, our method lacks access to dense, ground truth depth maps and full 3D supervision. Further, the 3D UNet architecture employed by NeSF is based on [26], a predecessor to the SparseConvNet architecture. As NeSF does not take advantage of sparsity, it must operate at a lower spatial resolution than the baseline and tends to mislabel small objects and thin structures. Though NeSF underperforms SparseConvNet today, we anticipate methodological improvements in model architecture to rapidly improve performance. Additional in-depth analysis is included in the [supplementary material](#).

Qualitative comparisons – Figure 5. Qualitatively, our method exhibits strong performance in identifying the 13 canonical categories in ToyBox13. Because our method operates directly on 3D geometry, it is not easily confused by objects of similar appearance but dissimilar geometry as demonstrated by the thin standing rifle in the top row. While

	ToyBox5				ToyBox13			
	TRAIN CAMERAS		TEST CAMERAS		TRAIN CAMERAS		TEST CAMERAS	
	2D mIoU	3D mIoU	2D mIoU	3D mIoU	2D mIoU	3D mIoU	2D mIoU	3D mIoU
NeSF	81.9 ± 0.8	88.7 ± 0.9	81.7 ± 0.6	89.6 ± 0.7	56.5 ± 0.8	60.1 ± 0.6	56.6 ± 1.0	61.9 ± 0.9
DeepLab [16]	81.6	N/A	N/A	N/A	63.1	N/A	N/A	N/A
SparseConvNet [44]	N/A	93.4	N/A	94.8	N/A	83.2	N/A	81.7

Table 3. **Quantitative comparison** – NeSF is competitive with 2D and 3D baselines. At train time, NeSF and DeepLab only utilize 2D supervision. Conversely, SparseConvNet requires full 3D supervision in the form of labeled 3D point clouds. We construct oracle point clouds via back-projected depth maps, resulting in an upper bound to our method (grayed-out row). Models are evaluated on train and test camera poses from test scenes. Configurations marked as “N/A” denotes a setting where methods are not applicable. Statistics for NeSF are aggregated across five random initializations.

Hyperparameter	2D	3D
Random Rotations	No	81.1
	Yes	92.0
Density Grid	(32, 32, 32)	87.5
	(48, 48, 48)	91.2
	(64, 64, 64)	91.7
	(80, 80, 80)	92.0
UNet	(16, 32, 64)	89.9
	(24, 48, 96)	91.5
	(32, 64, 128)	92.0
MLP	(0, 32)	91.3
	(1, 32)	91.8
	(1, 64)	91.2
	(2, 128)	92.0

Table 4. **Ablation: hyper-parameters** – Data augmentation, in the form of random scene rotations, increased spatial resolution of the density grid, and increased UNet model capacity improve 2D and 3D mIoU. Experiments on 25 scenes from the KLEVR dataset.

NeSF and SparseConvNet correctly recognize the rifle’s geometry, DeepLab labels it identically to the chair behind it. Similar to DeepLab, our method faces challenges with thin structures such as the tube of the standing lamp pictured in the middle row. Such structures are not well captured by the regular grids employed by DeepLab and NeSF in 2D and 3D, respectively. With access to dense, accurate point clouds, SparseConvNet correctly identifies the lamp in its entirety. One limitation particularly evident in NeSF is a tendency to smear labels across nearby objects, as demonstrated by the chair partially labeled as a display in the bottom row. Without access to *appearance* or fine-grained geometry, NeSF is unable to identify when one object ends and another begins. Integrating appearance information and access to higher spatial resolution are straightforward methods for improving NeSF’s accuracy in future work.

5.2. Ablation Studies

Our second set of experiments investigates how each component of our method affects system performance on the KLEVR dataset.

# RGB Images	NeRF		NeSF	
	PSNR	SSIM	2D	3D
5	21.2 ± 1.4	0.89 ± 0.02	52.2	72.4
10	24.2 ± 1.2	0.92 ± 0.01	79.6	96.7
25	30.3 ± 1.2	0.96 ± 0.01	87.3	97.0
50	35.5 ± 0.9	0.98 ± 0.00	90.8	97.3
75	37.5 ± 1.0	0.98 ± 0.00	91.4	97.1
100	38.4 ± 1.1	0.98 ± 0.00	92.0	97.4

Table 5. **Ablation: sensitivity to reconstruction quality** – The accuracy of our method improves with NeRF’s reconstruction quality. PSNR and SSIM are averaged across all scenes and metrics aggregated. Experiments on all scenes from KLEVR.

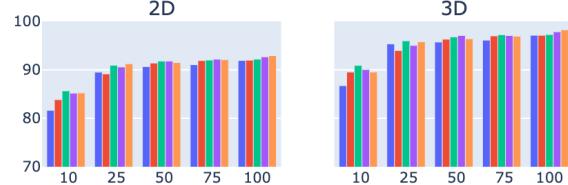


Figure 4. **Ablation: data efficiency** – 2D and 3D mIoU as a function of the number of train scenes for scenes with supervision from 1, 2, 5, 10, or 25 semantic maps per scene. NeSF generalizes to new scenes with as few as a *one semantic map per scene*. Additional semantic maps per scene marginally improve the accuracy. Experiments on KLEVR dataset.

Sensitivity to features. Table 4 shows results of an ablation study where one feature of our method is varied while holding all others to their reference values. We find that each component provides a measurable improvement in 2D and 3D segmentation quality. Data augmentation in the form of random scene rotations improves quality the most, adding 10.3% and 11.8% to 2D and 3D mIoU respectively. The spatial resolution of the probed NeRF density grids is the second most important as insufficient resolution makes smaller objects indistinguishable.

Sensitivity to reconstruction quality – Table 5. We investigate the robustness of NeSF to NeRF reconstruction quality. To modulate reconstruction quality, we vary the number of

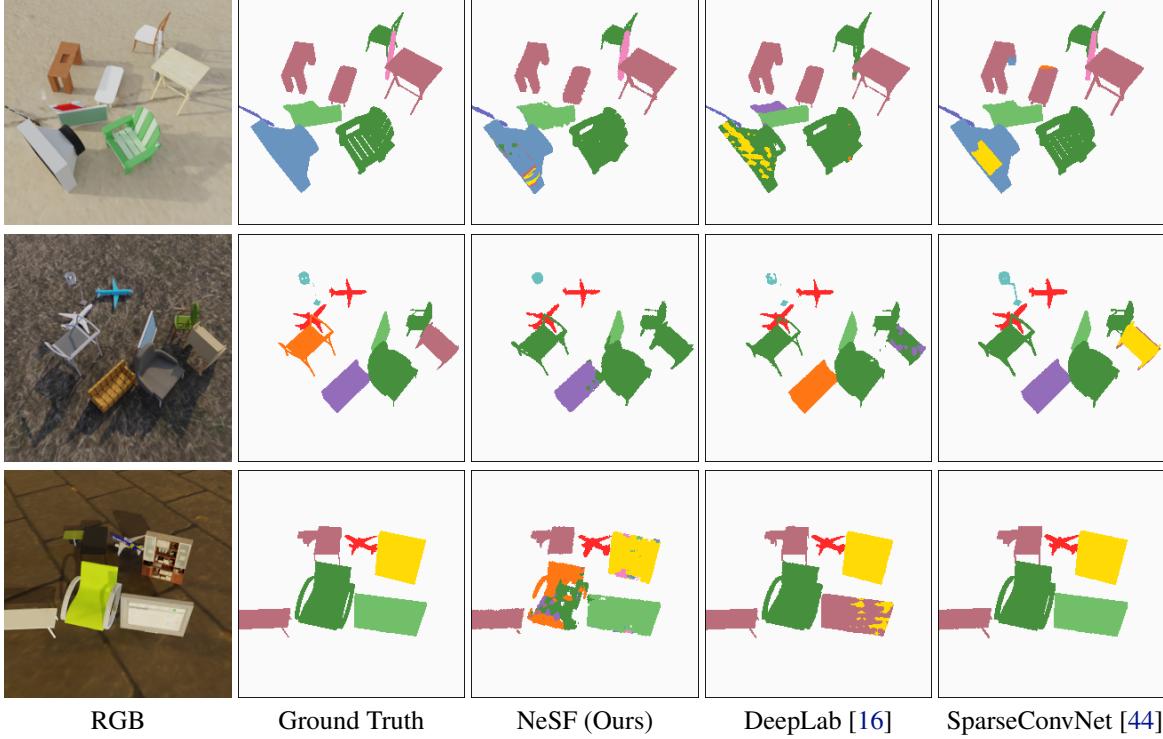


Figure 5. **Qualitative comparison (ToyBox13)** – Unlike DeepLab, NeSF is able to clearly separate objects with similar appearance but different geometry (Top). However, NeSF struggles with thin structures like lamp posts (Middle) and tends to smear labels from nearby objects (Bottom). SparseConvNet suffers from neither limitation but has access to oracle 3D geometry and full 3D supervision.

RGB images used when fitting NeRF models from 5 to 100. As expected, NeRF reconstruction quality as measured on novel views increases as more RGB images are provided. At the same time, we find that NeSF 2D segmentation accuracy improves monotonically with NeRF reconstruction quality. Improvement of NeRF reconstruction quality is a highly active area of research, as such we anticipate methodological improvements may be directly applied to improve NeSF’s performance. Surprisingly, 3D segmentation accuracy levels off near 97% when NeRF models are optimized with as little as 25 RGB images.

Sensitivity to data scarcity – Figure 4 To investigate NeSF’s applicability to scenarios where labeled semantic maps are scarce, we investigate robustness to the number of semantic maps per scene. We find NeSF easily generalizes to novel scenes with *as few as one semantic map per train scene*. Additional semantic maps per scene improve performance given a small number of scenes, with no noticeable effect after 25 scenes. This suggests datasets consisting of videos, each with a single labeled frame, are ideal for NeSF.

6. Conclusions and Limitations

In this work, we present NeSF, a novel method for simultaneous 3D scene reconstruction and semantic segmentation

from posed 2D images. Building on NeRF, our method is trained solely on posed 2D RGB images and semantic maps. At inference time, our method constructs a dense semantic segmentation field that can be queried directly in 3D or used to render 2D semantic maps from novel camera poses. We compare NeSF to competitive baselines in 2D and 3D semantic segmentation on three novel datasets.

In more challenging settings, we find that NeSF underperforms its baselines. However, NeSF offers novel capabilities. Unlike traditional 2D segmentation methods, NeSF fuses information across multiple independent views and renders semantic maps from novel poses. Unlike 3D point cloud methods, NeSF operates on posed 2D information alone at both train and test time. We chose to limit NeSF to a core set of features to better explore the fundamental trade-offs and capabilities of such an approach. We further explore trade offs with respect to model choice, and potential social impact in the [supplementary material](#). In future work, we anticipate extending NeSF to incorporate 2D semantic models and 3D sparsity will significantly improve accuracy.

In addition to NeSF, we propose three new datasets for multiview 3D reconstruction and semantic segmentation totalling over 3,000,000 frames and 1,000 scenes. Each dataset contains hundreds of scenes, each consisting of a set of randomly placed objects. The more challenging of these

datasets are rendered with realistic illumination and a large catalogue of objects and backgrounds. These datasets along with accompanying code and pretrained NeRF models, will be released to the public upon publication.

Acknowledgements

We would like to express our gratitude to Konstantinos Rematas, D. Sculley, and especially Thomas Funkhouser for their ideas and suggestions. We would like to note our deep appreciation for the project support and leadership given by Jakob Uszkoreit, without which the work would not have been possible. The authors would also like to thank Rocky Cai in his assistance in assembling the DeepLab baseline experiments.

References

- [1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33, 2020. [2](#)
- [2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. [2](#)
- [3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, pages 892–900, 2016. [2](#)
- [4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. [2](#)
- [5] Jonathan Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *ICCV*, October 2021. [3](#)
- [6] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446), 2019. [1](#)
- [7] Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. Visual navigation for mobile robots: A survey. *Journal of intelligent and robotic systems*, 53(3):263–296, 2008. [1](#)
- [8] Mark Boss, Raphael Braun, Varun Jampani, Jonathan Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural Reflectance Decomposition from Image Collections. In *ICCV*, October 2021. [3](#)
- [9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Lioung, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. [1](#)
- [10] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Lioung, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multi-modal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. [2, 5](#)
- [11] Rohan Chabra, Jan Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep Local Shapes: Learning Local SDF Priors for Detailed 3D Reconstruction. In *ECCV*, 2020. [3](#)
- [12] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. [2, 5](#)
- [13] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. [5](#)
- [14] Erzhuo Che, Jaehoon Jung, and Michael J Olsen. Object recognition, segmentation, and classification of mobile laser scanning point clouds: A state of the art review. *Sensors*, 19(4):810, 2019. [2](#)
- [15] Anpei Chen and Zexiang Xu. MVSNeRF: Fast Generalizable Radiance Field Reconstruction From Multi-View Stereo. In *ICCV*, October 2021. [3](#)
- [16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. [2, 6, 7, 8](#)
- [17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. [2](#)
- [18] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. BSP-Net: Generating Compact Meshes via Binary Space Partitioning. In *CVPR*, 2020. [3](#)
- [19] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. [2](#)
- [20] Zhiqin Chen and Hao Zhang. Learning Implicit Fields for Generative Shape Modeling. In *CVPR*, 2019. [3](#)
- [21] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12547–12556, 2021. [2](#)
- [22] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. [2](#)
- [23] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 1
- [24] 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. 2
- [25] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8958–8966, 2019. 2
- [26] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 4, 6, 15
- [27] Forrester Cole, Kyle Genova, Avneesh Sud, Daniel Vlasic, and Zhoutong Zhang. Differentiable surface rendering via non-differentiable sampling. In *ICCV*, pages 6088–6097, 2021. 3
- [28] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5
- [29] Tiago Cortinhal, George Tzelapis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. *arXiv preprint arXiv:2003.03653*, 2020. 2
- [30] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2, 5
- [31] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 1, 2
- [32] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 3
- [33] Boyang Deng, Kyle Genova, Sofien Bouaziz, Geoffrey Hinton, Andrea Tagliasacchi, and Soroosh Yazdani. CvxNet: Learnable Convex Decomposition. In *CVPR*, 2020. 3, 5
- [34] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: Neural Articulated Shape Approximation. In *ECCV*, 2020. 3
- [35] Terrance DeVries. Unconstrained Scene Generation With Locally Conditioned Radiance Fields. In *ICCV*, October 2021. 3
- [36] Bertrand Douillard, James Underwood, Noah Kuntz, Vsevolod Vlaskine, Alastair Quadros, Peter Morton, and Alon Frenkel. On the segmentation of 3d lidar point clouds. In *2011 IEEE International Conference on Robotics and Automation*, pages 2798–2805. IEEE, 2011. 1
- [37] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4), July 2018. 2
- [38] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7053–7062, 2019. 2
- [39] Stephan Garbin and Marek Kowalski. FastNeRF: High-Fidelity Neural Rendering at 200FPS. In *ICCV*, October 2021. 3
- [40] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William Freeman, and Thomas Funkhouser. Learning Shape Templates with Structured Implicit Functions. In *ICCV*, 2019. 3, 5
- [41] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. *3DV*, 2021. 1, 2, 16
- [42] Rohit Girdhar, Du Tran, Lorenzo Torresani, and Deva Ramanan. Distinit: Learning video representations without a single labeled video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 852–861, 2019. 2
- [43] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018. 2, 16
- [44] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 6, 7, 8
- [45] Klaus Greff and Andrea Tagliasacchi. Kubric. <https://github.com/google-research/kubric>, 2021. 5, 15
- [46] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 5
- [47] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2
- [48] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2
- [49] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, pages 1–19, 2020. 1
- [50] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2949, 2020. 2

- [51] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 1, 2
- [52] Yong He, Hongshan Yu, Xiaoyan Liu, Zhengeng Yang, Wei Sun, Yaonan Wang, Qiang Fu, Yanmei Zou, and Ajmal Mian. Deep learning based 3d segmentation: A survey. *arXiv preprint arXiv:2103.05423*, 2021. 2
- [53] Peter Hedman, Pratul Srinivasan, Ben Mildenhall, Jonathan Barron, and Paul Debevec. Baking Neural Radiance Fields for Real-Time View Synthesis. In *ICCV*, October 2021. 3
- [54] Philipp Henzler. Unsupervised Learning of 3D Object Categories from Videos in the Wild. In *CVPR*, 2021. 3
- [55] Alexander Hermans, Georgios Floros, and Bastian Leibe. Dense 3d semantic mapping of indoor scenes from rgb-d images. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2631–2638. IEEE, 2014. 2
- [56] Zeyu Hu, Mingmin Zhen, Xuyang Bai, Hongbo Fu, and Chiew-lan Tai. Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds. In *ECCV*, 2020. 2
- [57] Jingwei Huang, Haotian Zhang, Li Yi, Thomas Funkhouser, Matthias Nießner, and Leonidas J Guibas. Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4440–4449, 2019. 2
- [58] Rui Huang, Wanyue Zhang, Abhijit Kundu, Caroline Pantofaru, David A Ross, Thomas Funkhouser, and Alireza Fathi. An lstm approach to temporal 3d object detection in lidar point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 266–282. Springer, 2020. 16
- [59] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 1
- [60] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting Nerf on a Diet: Semantically Consistent Few-Shot View Synthesis. In *ICCV*, October 2021. 3
- [61] Wongbong Jang and Lourdes Agapito. CodeNeRF: Disentangled Neural Radiance Fields for Object Categories. In *ICCV*, October 2021. 3
- [62] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2
- [63] Yoonwoo Jeong. Self-Calibrating Neural Radiance Fields. In *ICCV*, October 2021. 3
- [64] Yiwei Jin, Diqiong Jiang, and Ming Cai. 3d reconstruction using deep learning: a survey. *Communications in Information and Systems*, 20(4):389–413, 2020. 2
- [65] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *arXiv preprint arXiv:2005.14169*, 2020. 2
- [66] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 5
- [67] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 15
- [68] A Sophia Koepke, Olivia Wiles, Yael Moses, and Andrew Zisserman. Sight to sound: An end-to-end approach for visual piano transcription. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1838–1842. IEEE, 2020. 2
- [69] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2
- [70] Adam Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Rezende. Nerf-VAE: A Geometry Aware 3D Scene Generative Model. In *ICML*, 2021. 3
- [71] Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jonathan Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, et al. The need 4 speed in real-time dense visual tracking. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018. 1
- [72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1
- [73] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *European Conference on Computer Vision*, pages 518–535. Springer, 2020. 1, 2
- [74] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3050–3057, 2014. 2
- [75] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, 338:321–348, 2019. 1
- [76] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3d semantic segmentation. In *International Conference on Computer Analysis of Images and Patterns*, pages 95–107. Springer, 2017. 2
- [77] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-Adjusting Neural Radiance Fields. In *ICCV*, October 2021. 3
- [78] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In

- European conference on computer vision*, pages 740–755. Springer, 2014. 1
- [79] David Lindell, Julien Martel, and Gordon Wetzstein. AutoInt: Automatic Integration for Fast Neural Volume Rendering. In *CVPR*, 2021. 3
- [80] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Sua, and Christian Theobalt. Neural Sparse Voxel Fields. In *Adv. Neural Inform. Process. Syst.*, 2020. 2, 3, 4
- [81] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128(2):261–318, 2020. 1
- [82] Weiping Liu, Jia Sun, Wanyi Li, Ting Hu, and Peng Wang. Deep learning on point clouds and its application: A survey. *Sensors*, 19(19):4188, 2019. 2
- [83] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2
- [84] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 598–605. IEEE, 2017. 2
- [85] Ricardo Martin-Brualla, Noha Radwan, Mehdi Sajjadi, Jonathan Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3
- [86] Ruben Mascaro, Lucas Teixeira, and Margarita Chli. Dif-fuser: Multi-view 2d-to-3d label diffusion for semantic scene segmentation. In *IEEE International Conference on Robotics and Automation (ICRA 2021)(virtual)*, 2021. 2
- [87] Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 37–45, 2015. 1
- [88] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3
- [89] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 2
- [90] Quan Meng. GNeRF: GAN-Based Neural Radiance Field Without Posed Camera. In *ICCV*, October 2021. 3
- [91] 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
- [92] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*, 2019. 3, 5
- [93] Johannes Meyer, Andreas Eitel, Thomas Brox, and Wolfram Burgard. Improving unimodal object recognition with multimodal contrastive learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. 2
- [94] Ben Mildenhall, Pratul Srinivasan, Matthew Tancik, Jonathan Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, pages 405–421. Springer, 2020. 2, 3, 4, 5, 15, 16
- [95] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss. RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2019. 2
- [96] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2
- [97] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, 2020. 2, 6
- [98] Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [99] Thomas Neff. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. In *Eurographics*, 2021. 3
- [100] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *CVPR*, 2021. 3
- [101] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *CVPR*, 2020. 3
- [102] Atsuhiro Noguchi. Neural Articulated Radiance Field. In *ICCV*, October 2021. 3
- [103] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *ICCV*, October 2021. 3
- [104] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2
- [105] 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
- [106] Jeong Joon Park, Pete Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *CVPR*, 2019. 3

- [107] Keunhong Park, Utkarsh Sinha, Jonathan Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. In *ICCV*, October 2021. 3
- [108] Sida Peng. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3
- [109] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision (ECCV)*, Cham, Aug. 2020. Springer International Publishing. 3
- [110] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. Jsis3d: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836, 2019. 2
- [111] Gerard Pons-Moll. Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion. In *CVPR*, 2020. 3
- [112] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2, 3, 4
- [113] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 2, 4
- [114] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. DeRF: Decomposed Radiance Fields. <https://arxiv.org/abs/2011.12490>, 2020. 2
- [115] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. KiloNeRF: Speeding Up Neural Radiance Fields With Thousands of Tiny MLPs. In *ICCV*, October 2021. 3
- [116] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. ShaRF: Shape-conditioned Radiance Fields from a Single View. In *ICML*, 2021. 3
- [117] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3577–3586, 2017. 2
- [118] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1
- [119] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Unet: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2
- [120] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*, October 2019. 3
- [121] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Adv. Neural Inform. Process. Syst.*, 2020. 3
- [122] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020. 2
- [123] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 2
- [124] Pratul Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan Barron. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In *CVPR*, 2021. 3
- [125] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2
- [126] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew Davison. iMAP: Implicit Mapping and Positioning in Real-Time. In *ICCV*, October 2021. 3
- [127] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 5
- [128] Towaki Takikawa. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *CVPR*, 2021. 2, 3
- [129] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6411–6420, 2019. 2
- [130] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2
- [131] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 1
- [132] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-Rigid Neural Radiance Fields: Reconstruction and Novel View Synthesis of a Deforming Scene from Monocular Video. In *ICCV*, October 2021. 3

- [133] Vibhav Vineet, Ondrej Miksik, Morten Lidegaard, Matthias Nießner, Stuart Golodetz, Victor A Prisacariu, Olaf Kähler, David W Murray, Shahram Izadi, Patrick Pérez, et al. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 75–82. IEEE, 2015. 2
- [134] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020. 2
- [135] Haiyan Wang, Xuejian Rong, Liang Yang, Shuihua Wang, and Yingli Tian. Towards weakly supervised semantic segmentation in 3D graph-structured point clouds of wild scenes. In *BMVC*, page 284, 2019. 2
- [136] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, 2021. 2, 3, 6
- [137] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF–: Neural Radiance Fields Without Known Camera Parameters. <https://arxiv.org/abs/2102.07064>, 2021. 3
- [138] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 6, 16
- [139] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronan Basri, and Yaron Lipman. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Adv. Neural Inform. Process. Syst.*, 2020. 3
- [140] Lin Yen-Chen, Pete Florence, Jonathan Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In *IROS*, 2021. 3
- [141] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for Real-Time Rendering of Neural Radiance Fields. In *ICCV*, October 2021. 3, 18
- [142] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*, 2021. 2, 3, 6
- [143] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. STaR: Self-supervised Tracking and Reconstruction of Rigid Objects in Motion with Neural Rendering. In *CVPR*, 2021. 3
- [144] Greg Zaäl, Rob Tuytel, Rico Cilliers, James Ray Cock, Andreas Mischok, Sergej Majboroda, Dimitrios Savva, and Jutta Burger. Hdri haven. <https://polyhaven.com/hdris>, 2021. 5
- [145] Cheng Zhang, Zhi Liu, Guangwen Liu, and Dandan Huang. Large-scale 3d semantic mapping using monocular vision. In *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pages 71–76. IEEE, 2019. 2
- [146] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Ner-factor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970*, 2021. 2
- [147] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020. 2
- [148] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 570–586, 2018. 2
- [149] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *ICCV*, October 2021. 2, 3, 4
- [150] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9939–9948, 2021. 2

NeSF: Neural Semantic Fields for Generalizable Semantic Segmentation of 3D Scenes

Supplementary Material

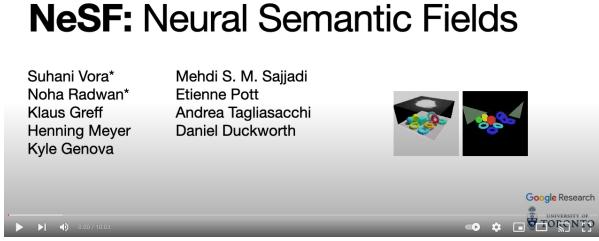


Figure 6. **Overview Video** – We highly recommend viewing the high-definition overview video on [YouTube](#).

A. Contributions

We describe the specific contributions of each individual author in the following.

Suhani Vora was one of two first authors on this work. She implemented NeSF’s 2D inference framework; implemented and ran NeSF experiments; implemented and ran DeepLab baseline experiments; setup the project website; and wrote portions of the paper.

Noha Radwan was one of two first authors on this work. She implemented the 3D semantic reasoning module; implemented and ran NeSF experiments; implemented NeSF’s 2D evaluation framework; setup the project website; and edited the paper.

Klaus Greff was responsible for Kubric [45], the technology used to generate the datasets in this work. He further provided assistance in development of the datasets and the animations in the overview video.

Henning Meyer assisted in the development of the ToyBox5 and ToyBox13 datasets. He further contributed significantly to the codebase upon which NeSF was built.

Kyle Genova was responsible for the SparseConvNet baseline. He further assisted in experiment design and wrote portions of the paper.

Mehdi S. M. Sajjadi suggested experiments and contributed to the codebase upon which NeSF was built

Etienne Pott implemented the scalable dataset pipeline used to train NeSF. He further contributed significantly to the codebase upon which NeSF was built.

Andrea Tagliasacchi oversaw the project as a research lead. He proposed the initial NeSF model architecture He further

suggested experiments; and oversaw and wrote large portions of the paper.

Daniel Duckworth oversaw the project as a technical lead. He coordinated contributors and designed and prototyped the software architecture within which NeSF was implemented. He further suggested, implemented, and ran NeSF experiments; generated the datasets used in this paper; implemented NeSF’s 3D inference framework; wrote portions of the paper; generated visualizations; and wrote, assembled, and recorded the overview video.

B. Training Details

We describe the model architecture and training procedure used by NeSF and its baselines below. Unless otherwise stated, we train each method using all TRAIN SCENES with 9 randomly-selected images per scene. For KLEVR, this results in 100 TRAIN SCENES; for ToyBox5 and 500 TRAIN SCENES for ToyBox13. All methods are evaluated on 4 randomly-selected images from each dataset’s NOVEL SCENES. For KLEVR, this results in 20 NOVEL SCENES; for ToyBox5 and ToyBox13, 25 NOVEL SCENES each. We ensure that each method observes the same randomly-selected set of images per scene by specifying the seed of the random number generator.

NeRF. The first stage of NeSF is the training of per-scene NeRF models. We employ the model architecture and training regime of Mildenhall *et al.* [94]. Each scene’s density field is described by an MLP with 8 hidden layers of 256 units, and its appearance by an additional MLP of 1 hidden layer and 128 units. We employ 10 octaves for positional encoding. Each NeRF model is trained on pixels selected at random from 9 views with the Adam optimizer [67]. The learning rate is exponentially decayed from $1e-3$ to $5.4e-4$ over 25,000 steps. We train each NeRF model for approximately 20 minutes on 8 TPUs v2 cores.

NeSF. NeSF has two major model components: a 3D UNet and an MLP Decoder. For the 3D UNet, we employ the UNet architecture of Çiçek *et al.* [26] with the BatchNorm layers removed and only 2 max-pooling operations. We use 32, 64, and 128 output channels prior to each max-pooling operation. For the MLP Decoder, we employ 2 hidden layers of 128 hidden units each with a ReLU non-linearity.

We train NeSF with Adam optimizer [67]. We use an exponentially decaying learning rate initialized to $1e-3$ and decaying to $1e-5$ over 25,000 steps. At each step, we employ a stratified sampling approach: we randomly select 32

scenes, then randomly select a set of 128 pixels from each scene’s TRAIN CAMERAS. For volumetric rendering, we sample 192 points along each ray according to the stratified approach employed in NeRF [94]. For each scene in the batch, we discretize NeRF’s density grid by probing at 64^3 evenly-spaced points. Before discretizing, we apply a random rotation about the z-axis (upwards) to each scene. For smoothness regularization, we uniformly sample 8,192 additional 3D coordinates from each scene and add random noise with standard deviation 0.05. When computing the loss, we assign a weight of 0.1 to the smoothness regularization term.

We find that we are able to train NeSF to convergence in approximately 45 minutes on 32 TPUv3 cores.

DeepLab. We train a DeepLab Wide-ResNet-38 model [138], warm starting with a checkpoint pre-trained on COCO. For our optimization scheme, we apply SGD + Momentum with a slow start learning rate of 1e-4 and a linear ramp up to 6e-3 followed by a cosine schedule decay in learning rate to 1.26e-7 at 55,000 training steps. We additionally apply weight decay of $1.0e-4$. For each train step, we use a batch size of 32. Models are trained on 32 TPUv3 chips. To enable re-use of a well-performing hyperparameter configuration, we up-sample our input images from 256x256 to 1024x1024, using bilinear interpolation for the RGB input and nearest neighbor interpolation for the corresponding semantic maps.

SparseConvNet. Our SparseConvNet [43] implementation is based on the TF3D [58] and 2D3DNet [41] implementations. Each convolutional layer except the last is an occupancy-normalized $3 \times 3 \times 3$ sparse spatial convolution followed by batch norm and then ReLU. The final layer omits batch norm and ReLU. Each encoder stage is a pair of convolution layers followed by a $2 \times 2 \times 2$ spatial max-pool operation, and each decoder layer is a voxel unpooling operation followed by a pair of convolutional layers. The encoder feature widths are $(64, 64), (64, 96), (96, 128), (128, 160), (160, 192), (192, 224), (224, 256)$. These are the output channel counts of the first and second convolutional layers per block. The bottleneck is a sequence of two convolutional layers of widths 256 each. The decoder feature widths are $(256, 256), (224, 224), (192, 192), (160, 160), (128, 128), (96, 96), (64, 64)$. Finally we apply a sequence of three convolutional layers with sizes $(64, 64, \text{class_count})$, which are followed by a softmax layer and a cross-entropy loss function. Our input features are only occupancy (i.e., a 1 on all input points). We use 0.005-width voxels in a $[-1, 1]$ cube scene. We optimize for 450,000 steps with SGD using a momentum of 0.9, a batch size of 5, an initial learning rate of 0.015, and a cosine learning rate decay starting at step 200,000 and ending at step 450,000. We add an ℓ_2 weight decay loss of $1e-4$ and train asynchronously on 20 NVIDIA V100 GPUs. We apply the following data augmentations: XY rotations of up to ± 10 degrees, z rotations of ± 180 degrees, and a random scale factor between 0.9 and 1.1.

C. Analysis

C.1. Qualitative Results

In Figures 12, 13, and 14, we present randomly-selected qualitative results on each dataset studied in this paper. In each row, we depict the ground truth RGB, depth, and semantic map alongside 2D segmentation maps produced by NeSF, DeepLab, and SparseConvNet. We observe that all methods are effective at separating foreground objects from the floor and background. Unlike SparseConvNet, NeSF and DeepLab tend to assign different parts of the same object to different semantic categories when the correct category is ambiguous.

While NeSF and SparseConvNet are multi-view consistent by design, this is not the case for 2D methods such as DeepLab. In Figure 7, we demonstrate one instance of 3D inconsistency. In this example, NeSF and SparseConvNet label the orange couch and white-blue display identically from both views, whereas DeepLab’s classification changes.

NeSF 3D Density Field Quality. Notably, a key difference between SparseConvNet and NeSF is the provision of a ground truth point cloud as input for SparseConvNet. Several aspects of SparseConvNet may contribute to its overall superior performance relative to NeSF including access to oracle 3D geometry, sparse point cloud input representation, or the SparseConvNet model architecture. To better understand where headroom exists for improvement of NeSF, we begin by visually inspecting the difference in 3D geometry between the 3D density field of NeSF in relation to the ground truth point cloud provided to SparseConvNet for scenes selected from KLEVR, ToyBox5, and ToyBox13 in Figure 8. We observe the NeSF density fields often miss thin structures and fine details, and “floaters” are particularly evident in ToyBox13. Improvement of density field quality via improvements to NeRF representations may resolve such artifacts. Furthermore, akin to results in Table 7, such improvements would likely improve the performance of NeSF. We leave additional inspection of 3D input representation and replacement of the semantic model architecture of NeSF for future work.

C.2. Ablations

Model Ablations – Table 6. We repeat our ablation study on the ToyBox5 model and observe overall consistent results with KLEVR model ablations. Table 6 shows results varying each component. Similar to the results on KLEVR, we observe that data augmentation in the form of random scene rotations improves quality the most, adding 9.3% and 6.1% to 2D and 3D mIoU respectively. The spatial resolution of the probed NeRF density grids is again confirmed as crucial, and notably to a greater extent than for KLEVR. We hypothesize that this occurs as ToyBox5 contains more fine structured objects than KLEVR.

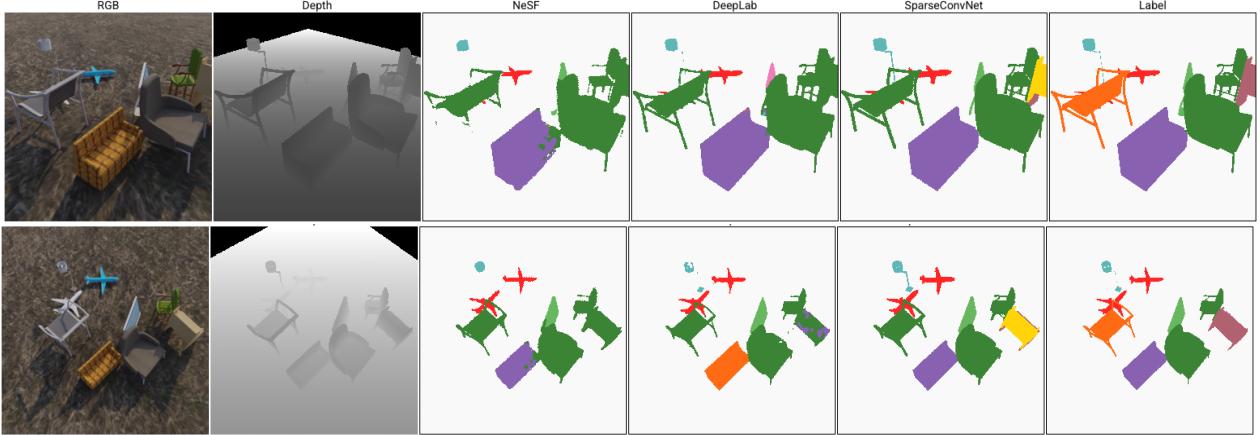


Figure 7. **Multiview Consistency** While NeSF and SparseConvNet classify the orange couch and the display identically from multiple independent views of the same scene, DeepLab’s predictions vary.

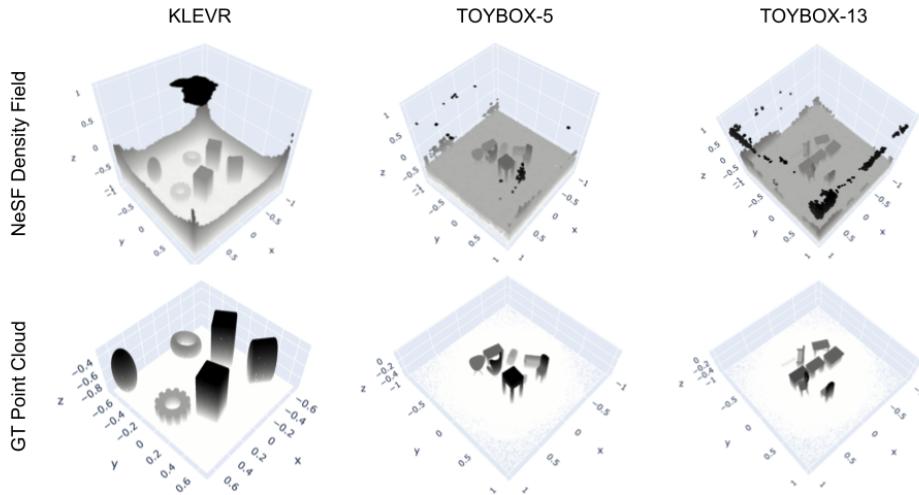


Figure 8. **NeRF 3D Density Field (top)** and **Ground Truth Point Clouds (bottom)** for **KLEVR**, **ToyBox5**, and **ToyBox13**. We find that NeRF’s density field accurately captures the 3D geometry of the scene. The NeSF density fields are sampled at a resolution of 128x128x128, and are filtered for ease of visualization for positive σ values with thresholds of 16, 64, and 64 for the 3 datasets respectively.

Sensitivity to reconstruction quality – Table 7. We re-evaluate the robustness of NeSF to NeRF reconstruction quality in the context of the ToyBox5 dataset. To modulate reconstruction quality, we vary the number of RGB images used when fitting NeRF models from 5 to 100 and confirm NeRF reconstruction quality improves as more RGB images are provided. As previously observed, the 2D and 3D segmentation quality of NeSF improves monotonically with NeRF’s reconstruction quality. Notably, 3D segmentation accuracy begins to level off near 88% when NeRF models are optimized with as few as 50 RGB images, with a large jump in performance between 25 and 50 images.

Sensitivity to data scarcity – Figure 9 We repeat our analysis providing limited numbers of semantically labelled

maps for NeSF’s training on the ToyBox5 model. We vary the number of provided label maps from 1 to 50. Similar to the KLEVR setup, we observe that providing additional semantic maps per scene improves the performance, with a large jump between 5 and 10 maps. A saturation in the model performance is reached at around 25 maps per scene. Moreover, the model is still able to generalize with as little as 1 semantic map per scene.

C.3. Multiview Consistency

Unlike conventional 2D methods, NeSF is 3D-consistent by design. In Figure 10, we visualize the epipolar plane traced out along the red reference line for NeSF’s semantic predictions. We find that the resulting predictions are

Hyperparameter		2D	3D
Random Rotations	No	69.5	83.6
	Yes	78.8	89.7
Density Grid	(32, 32, 32)	71.1	81.5
	(48, 48, 48)	76.4	89.3
	(64, 64, 64)	78.8	89.7
UNet	(16, 32, 64)	80.6	89.1
	(24, 48, 96)	80.1	89.8
	(32, 64, 128)	79.0	89.8
MLP	(0, 32)	78.6	90.7
	(1, 32)	79.7	89.8
	(1, 64)	80.7	89.4
	(2, 128)	79.2	89.5

Table 6. **Ablation: hyper-parameters** – Data augmentation, in the form of random scene rotations, increased spatial resolution of the density grid, and increased UNet model capacity improve 2D and 3D mIoU. Experiments on 500 scenes from the ToyBox5 dataset.

# RGB Images	NeRF		NeSF	
	PSNR	SSIM	2D	3D
5	17.5 ± 2.1	0.55 ± 0.15	15.0	17.9
10	19.2 ± 2.9	0.62 ± 0.15	29.1	35.2
25	23.9 ± 2.7	0.76 ± 0.09	61.4	74.1
50	26.3 ± 2.1	0.81 ± 0.06	72.3	88.7
75	27.3 ± 2.0	0.83 ± 0.05	72.6	89.5
100	27.9 ± 2.0	0.84 ± 0.04	73.6	90.0

Table 7. **Ablation: sensitivity to reconstruction quality for ToyBox5** – The accuracy of our method improves with NeRF’s reconstruction quality. PSNR and SSIM are averaged across all scenes and metrics aggregated. Experiments on all scenes from ToyBox5.

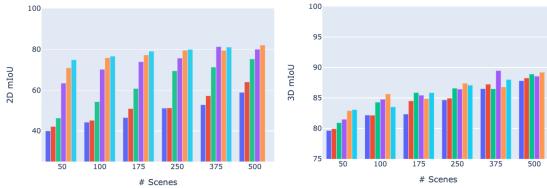


Figure 9. **Ablation: data efficiency** – 2D and 3D mIoU as a function of the number of train scenes for scenes with supervision from 1, 2, 5, 10, 25, or 50 semantic maps per scene. Note, 50 semantic maps did not fit in memory for 500 scenes, hence this particular setup is dropped from the results. NeSF generalizes to new scenes with as few as a *one semantic map per scene*. Additional semantic maps per scene marginally improve the accuracy. Experiments on ToyBox5 dataset.

consistent and smooth except when a floater obstructs the camera’s view, as illustrated by the white smudge in the EPI. In Figure 11, we see a further example of this phenomena. As the camera rotates about the scene, a floating mass of density obstructs the camera’s view, and the resulting seman-

tic maps contain a large number of mislabeled pixels. **We strongly encourage the reader to view additional results in the accompanying video for more detail.**

C.4. Limitations

Confusion matrix. In Table 8 and Table 9, we present NeSF’s per-class confusion matrix on the ToyBox13 dataset for 2D pixel and 3D point classification. While NeSF is able to easily identify larger, articulated semantic categories such as cabinet, chair, display, or table (78.0-89.4% 2D, 79.4-93.6% 3D), it struggles to identify object categories for smaller objects such as rifle (56.3% 2D, 75.3% 3D) or geometrically unarticulated objects such as loudspeaker (38.5% 2D, 40.4% 3D). When NeSF confuses foreground object categories, the most common errors are between geometrically-similar classes. For example, benches are often mislabeled as chairs (17.0% 2D, 17.9% 3D) and sofas (26.5% 2D, 27.4% 3D), and loudspeakers are often mislabeled as tables (32.0% 2D, 32.5% 3D).

Accuracy 2D vs. 3D. Our experiments indicate that NeSF’s accuracy is *higher* in 3D than in 2D. We found this surprising, especially considering the 2D nature of NeSF’s semantic supervision. We believe the ultimate cause to be “floaters” in the 3D density field recovered by NeRF. In Table 8, we see that approximately 5-10% of 2D pixels from each semantic category are mislabeled as “background”. In contrast, Table 9 demonstrates that the same type of error is made approximately 1% of the time in 3D. The most prominent exception to this is the bench category, whose objects often contain thin structures poorly captured by NeSF.

Impact of floaters. In Figure 11, we qualitatively show how “floaters” reduce NeSF’s accuracy in image-space. In this set of 5 video frames, we demonstrate a camera path passing in front of a floating cloud. This cloud is assigned to the background semantic category and obscures the foreground objects from the scene. In spite of the *objects* being correctly labeled, the generated semantic maps are incorrect. As a result, NeSF achieves lower 2D mIoU than 3D mIoU as the latter is not hindered by floaters and is corroborated by Tables 8 and 9. We believe that eliminating this failure in geometric construction will significantly improve NeSF’s accuracy. Solutions are readily provided by methods building on NeRF [141]. **We strongly encourage the reader to view additional results in the accompanying video for more detail.**

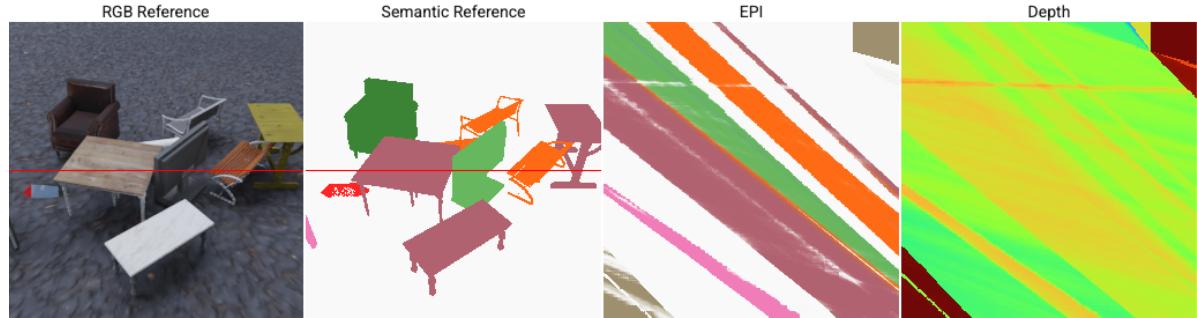


Figure 10. **Epipolar Plane** We demonstrate the 3D consistency of NeSF by rendering the epipolar plane along the red scan line as the camera moves from right to left. The epipolar plane is smooth and consistent except when a “floater” passes in front of the camera.

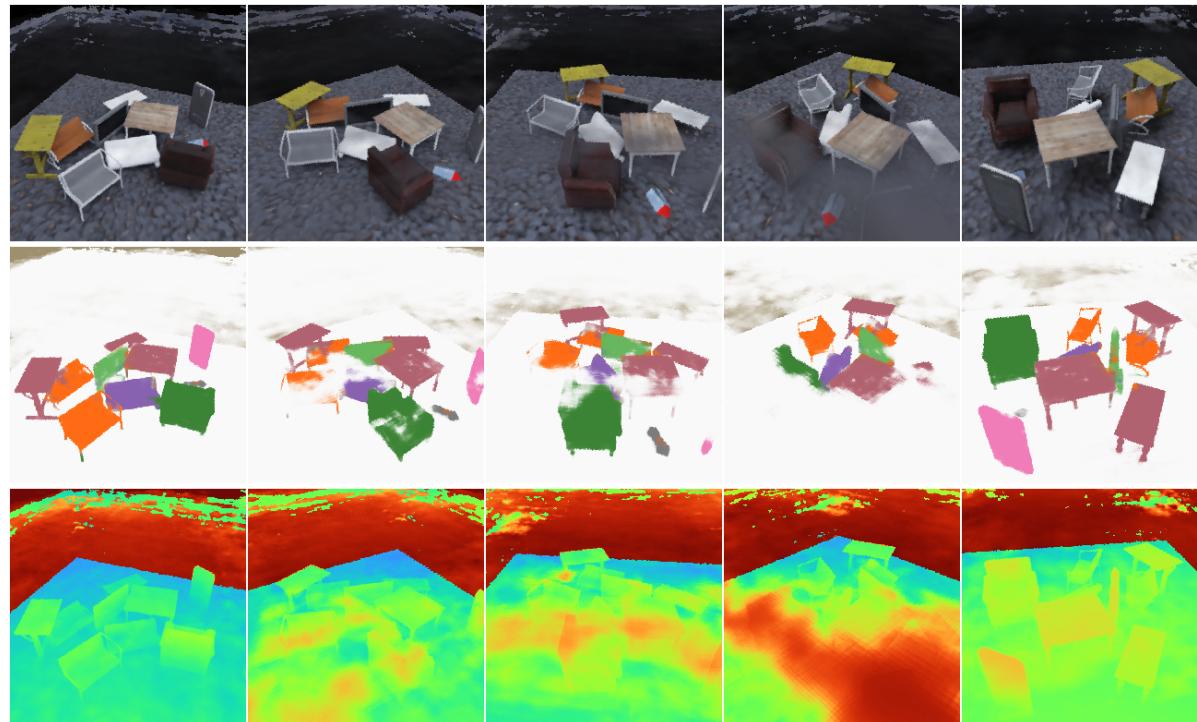


Figure 11. **Floating Objects** In the above 3 rows, we illustrate NeRF’s RGB reconstruction (top), NeSF’s semantic field, and NeRF’s density field (bottom) over the course of 5 video frames. When NeRF’s density field contains “floaters”, NeSF often assigns them to the background semantic category.

	background	airplane	bench	cabinet	car	chair	display	lamp	loudspeaker	rifle	sofa	table	telephone	vessel
background	99.1%	0.1%	0.0%	0.0%	0.2%	0.1%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%
airplane	12.9%	69.2%	3.2%	0.1%	0.0%	1.9%	0.0%	0.3%	0.0%	1.4%	0.0%	4.6%	0.0%	6.3%
bench	0.1%	0.7%	42.1%	0.1%	0.0%	17.0%	0.0%	0.2%	0.0%	1.2%	26.5%	1.7%	0.0%	1.4%
cabinet	2.2%	0.0%	0.0%	78.0%	0.0%	0.1%	2.7%	0.0%	12.8%	0.0%	0.0%	2.7%	1.5%	0.0%
car	5.0%	0.1%	0.1%	0.5%	84.6%	0.0%	0.1%	0.3%	0.1%	0.2%	1.3%	4.1%	0.0%	3.3%
chair	5.8%	0.0%	0.0%	0.0%	0.0%	89.4%	0.1%	0.0%	0.0%	0.1%	0.0%	2.9%	0.7%	0.0%
display	9.2%	0.0%	0.3%	0.4%	0.1%	1.2%	83.3%	0.0%	3.7%	0.0%	0.2%	1.3%	0.2%	0.0%
lamp	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%	61.4%	0.0%	18.1%	0.0%	0.0%	0.1%	7.8%	0.8%
loudspeaker	5.4%	0.1%	4.3%	10.5%	0.0%	0.6%	3.2%	1.1%	38.5%	0.0%	2.2%	32.0%	1.7%	0.4%
rifle	30.7%	1.6%	2.0%	0.2%	0.6%	0.7%	3.1%	0.0%	0.0%	50.3%	3.0%	0.7%	0.0%	1.7%
sofa	10.0%	2.3%	0.6%	0.0%	0.0%	21.6%	0.0%	0.1%	0.1%	0.0%	61.4%	0.0%	0.0%	3.6%
table	5.3%	0.0%	1.7%	3.1%	0.0%	4.6%	0.1%	3.1%	3.1%	0.0%	0.7%	79.9%	0.0%	0.3%
telephone	1.7%	0.0%	0.0%	9.3%	0.0%	0.2%	3.7%	0.0%	15.7%	0.0%	0.0%	0.3%	69.0%	0.0%
vessel	5.9%	12.4%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	4.3%	0.3%	0.0%	0.0%	77.0%

Table 8. **Confusion matrix for 2D semantic segmentations by NeSF on ToyBox13.** Each row corresponds to a ground truth label and is normalized to sum to 100%. NeSF’s most common errors include confusing similarly-shaped objects and classifying small and thin objects as background. Correct classifications are highlighted in bold.

	background	airplane	bench	cabinet	car	chair	display	lamp	loudspeaker	rifle	sofa	table	telephone	vessel
background	99.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
airplane	1.7%	78.0%	3.6%	42.7%	0.2%	0.0%	17.9%	0.0%	0.2%	0.0%	2.1%	4.8%	0.1%	7.6%
bench	6.0%	0.8%	0.0%	79.4%	0.0%	0.0%	2.9%	0.0%	0.3%	0.0%	1.7%	27.4%	1.6%	1.4%
cabinet	0.4%	0.0%	0.3%	86.7%	0.6%	0.0%	0.0%	0.0%	0.2%	0.4%	0.3%	0.0%	2.6%	0.0%
car	0.8%	0.4%	0.0%	1.1%	93.6%	0.0%	0.1%	0.1%	0.1%	0.1%	3.1%	4.0%	0.0%	4.3%
chair	1.0%	0.0%	0.1%	0.4%	0.3%	1.0%	0.1%	0.1%	0.1%	0.1%	0.8%	0.0%	0.0%	0.0%
display	0.6%	0.0%	0.0%	0.1%	0.1%	0.0%	66.7%	0.0%	0.3%	4.8%	0.0%	0.4%	0.3%	0.0%
lamp	2.3%	0.0%	0.0%	12.5%	0.1%	0.1%	0.4%	0.0%	20.7%	0.0%	0.0%	0.1%	8.8%	0.9%
loudspeaker	1.1%	0.1%	4.4%	4.6%	0.6%	0.1%	1.8%	3.6%	0.0%	40.4%	0.0%	2.1%	32.5%	1.9%
rifle	2.4%	4.9%	2.9%	0.8%	0.0%	0.0%	22.8%	0.0%	0.1%	0.1%	0.0%	3.2%	0.9%	0.0%
sofa	0.7%	0.0%	1.8%	3.5%	0.0%	4.7%	0.0%	3.2%	1.5%	68.2%	0.0%	0.0%	4.3%	0.0%
table	0.6%	0.1%	0.0%	10.5%	0.0%	0.0%	4.5%	0.0%	15.8%	0.0%	0.8%	82.8%	0.0%	0.8%
telephone	0.1%	1.1%	13.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.4%	0.6%	0.0%	60.1%	0.0%
vessel														80.7%

Table 9. **Confusion matrix for 3D semantic segmentations by NeSF on ToyBox13.** Each row corresponds to a ground truth label and is normalized to sum to 100%. NeSF’s most common errors include confusing similarly-shaped objects and classifying small and thin objects as background. Correct classifications are highlighted in bold.

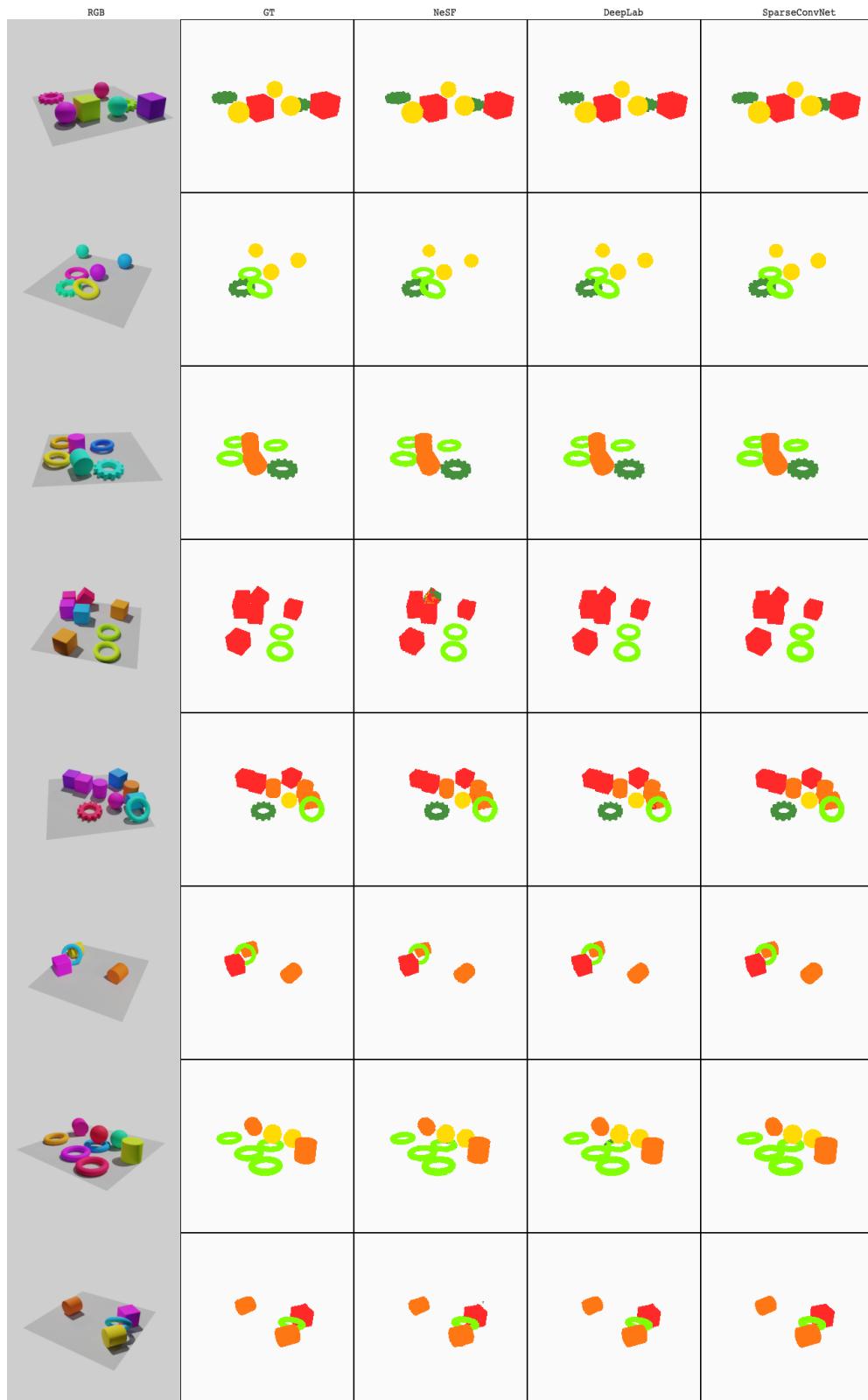


Figure 12. Additional Qualitative Results on KLEVR

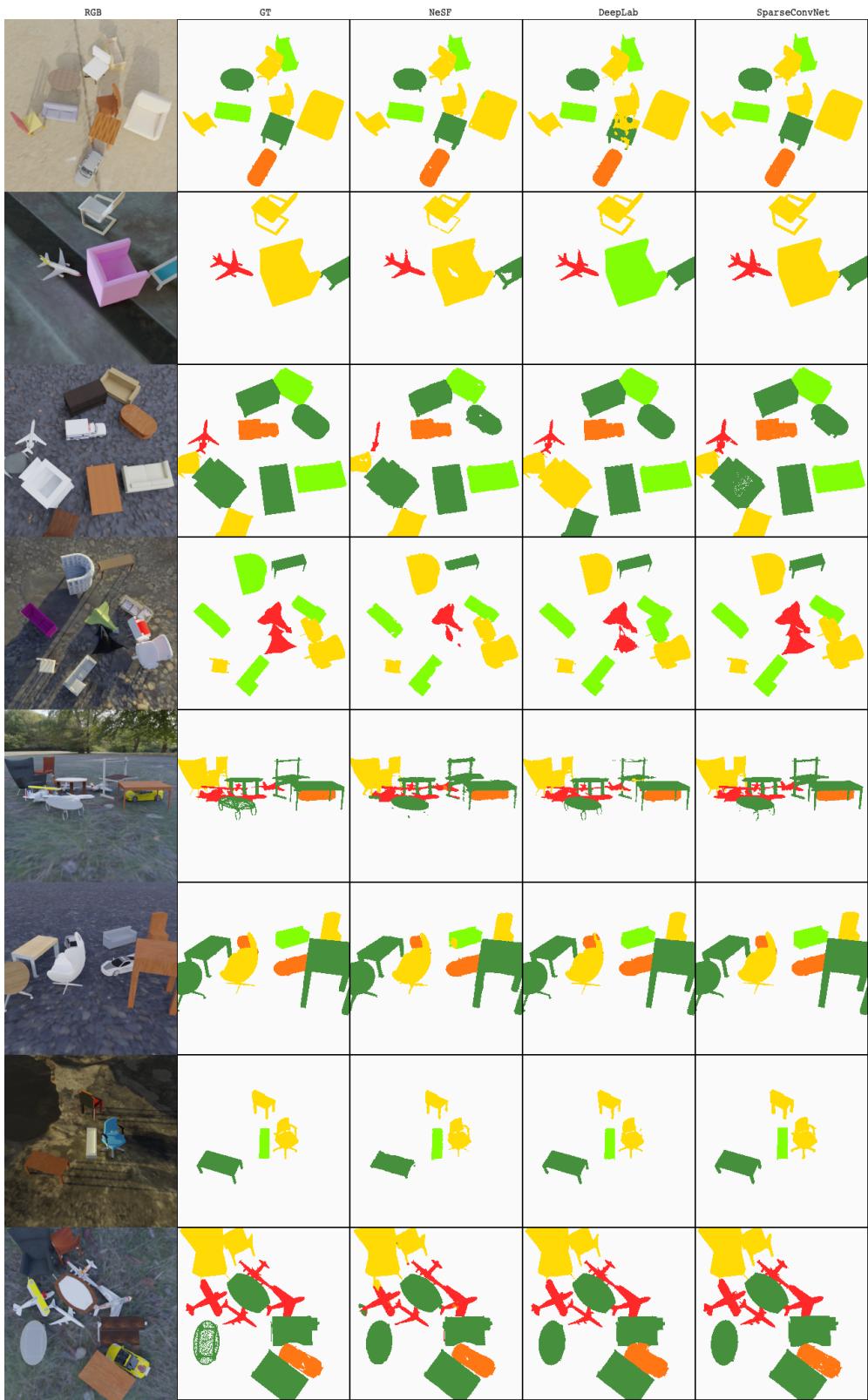


Figure 13. Additional Qualitative Results on ToyBox5

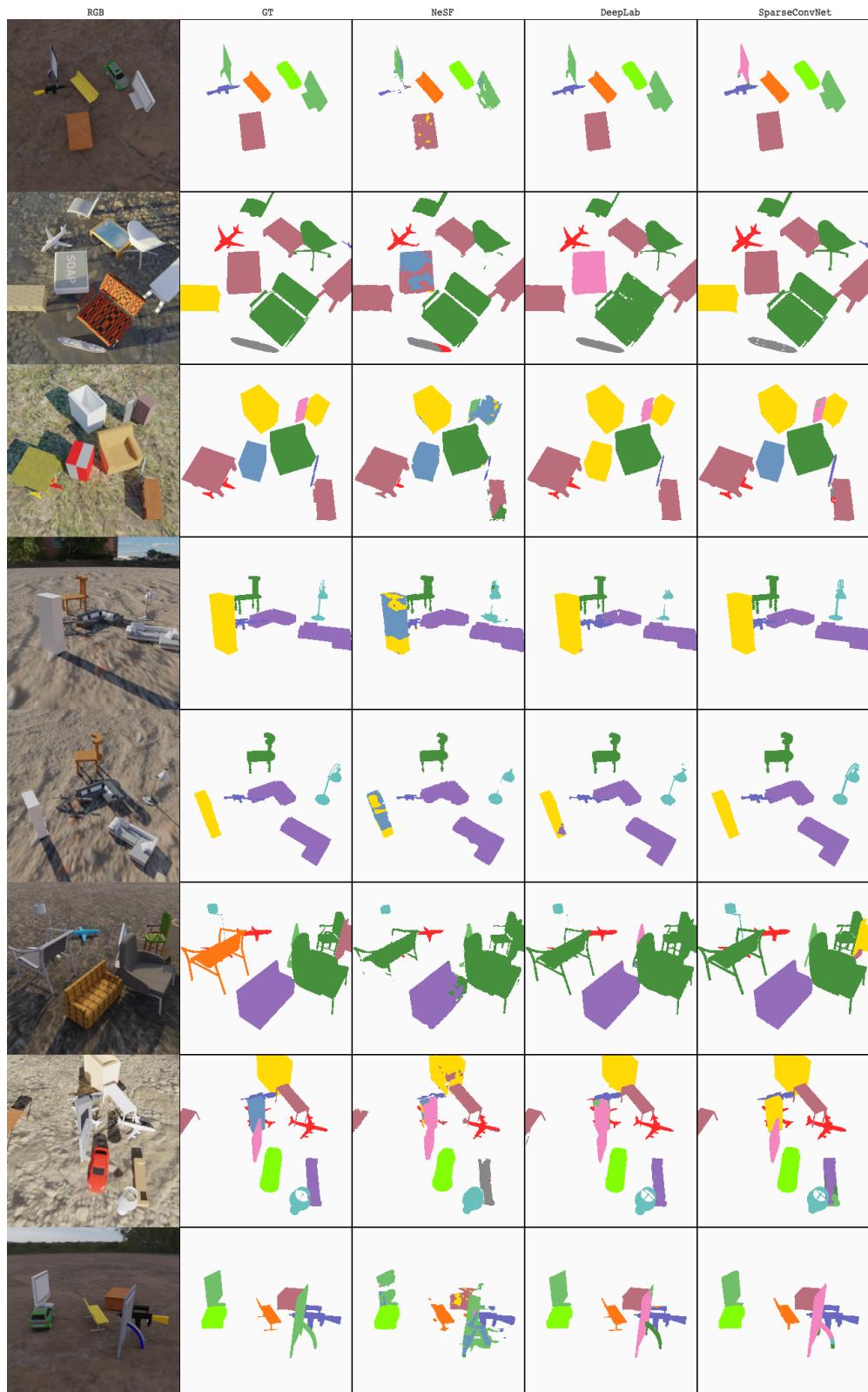


Figure 14. Additional Qualitative Results on ToyBox13