NeO 360: Neural Fields for Sparse View Synthesis of Outdoor Scenes

Muhammad Zubair Irshad[1,2] Sergey Zakharov[2] Katherine Liu[2] Vitor Guizilini[2] Thomas Kollar[2]
Adrien Gaidon[2] Zsolt Kira*[1] Rares Ambrus*[2]

* denotes shared last authorship

[1]Georgia Institute of Technology [2]Toyota Research Institute

{mirshad7, zkira}@gatech.edu, {firstname.lastname}@tri.global

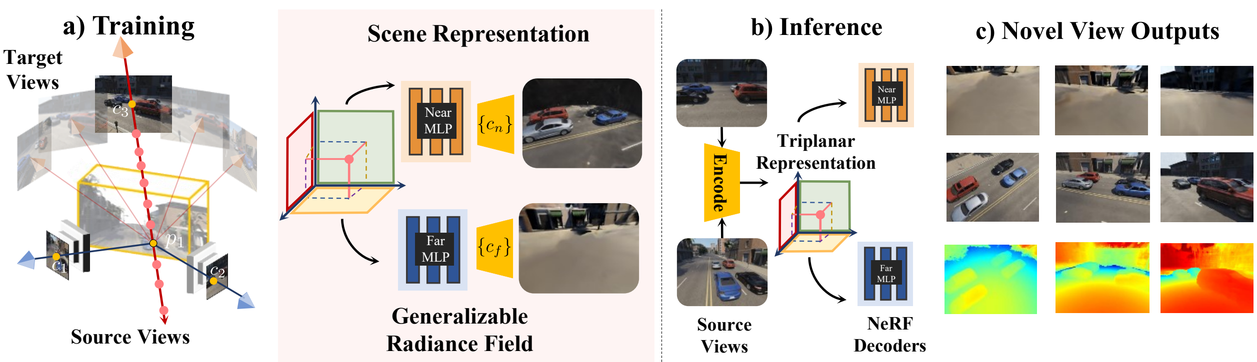

Figure 1: **Overview:** a) Given just a single or a few input images from a novel scene, our method renders new 360° views of complex unbounded outdoor scenes b) We achieve this by constructing an image-conditional triplane representation to model the 3D surrounding from various perspectives. c) Our method *generalizes* across novel scenes and viewpoints for complex 360° outdoor scenes.

Abstract

Recent implicit neural representations have shown great results for novel view synthesis. However, existing methods require expensive per-scene optimization from many views hence limiting their application to real-world unbounded urban settings where the objects of interest or backgrounds are observed from very few views. To mitigate this challenge, we introduce a new approach called NeO 360, Neural fields for sparse view synthesis of outdoor scenes. NeO 360 is a generalizable method that reconstructs 360° scenes from a single or a few posed RGB images. The essence of our approach is in capturing the distribution of complex real-world outdoor 3D scenes and using a hybrid image-conditional triplanar representation that can be queried from any world point. Our representation combines the best of both voxel-based and bird's-eye-view (BEV) representations and is more effective and expressive than each. NeO 360's representation allows us to learn from a large collection of unbounded 3D scenes while offering generalizability to new views and novel scenes from as few as a single image during inference. We demonstrate our approach on the proposed challenging 360° unbounded dataset, called NeRDS 360, and show that NeO 360 outperforms state-of-the-art generalizable methods for novel view synthesis while also offering editing and composition capabilities. Project page: zubair-irshad.github.io/projects/neo360.html

1. Introduction

Advances in learning-based implicit neural representations have demonstrated promising results for high-fidelity novel-view synthesis [64], doing so from multi-view images [60, 62]. The capability to infer accurate 3D scene representations has benefits in autonomous driving [42, 16] and robotics [22, 41, 21].

Despite great progress in neural fields [57] for indoor novel view synthesis, these techniques are limited in their ability to represent complex urban scenes as well as decompose scenes for reconstruction and editing. Specifically, previous formulations [6, 46, 73] have focused on per-scene optimization from a large number of views, thus increasing their computational complexity. This requirement limits their application to complex scenarios such as data cap-

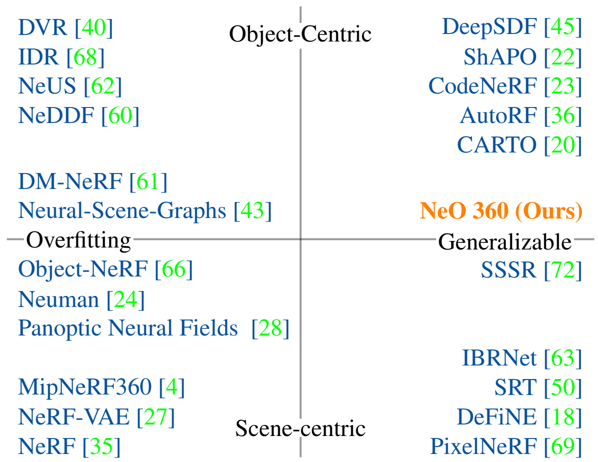

Figure 2: **Taxonomy of implicit representation methods** reconstructing appearance and shapes. The x and y correspond to (Generalizable vs Overfitting) and (Object-Centric vs Scene-Centric) dimensions, as discussed in Section 1.

tured by a moving vehicle where the geometry of interest is observed in just a few views. Another line of work focuses on object reconstructions [36, 22] from single-view RGB (Fig. 2). However, these approaches require accurate panoptic segmentation and 3D bounding boxes as input which is a strong supervisory signal and consists of multi-stage pipelines that can lead to error-compounding.

To avoid the challenge of acquiring denser input views of a novel scene in order to obtain an accurate 3D representation as well as the computational expense of per-scene optimization from many views, we propose to infer the representation of 360° unbounded scenes from just a single or a few posed RGB images of a novel outdoor environment.

As shown in Fig. 1, our approach extends the NeRF++ [73] formulation by making it generalizable. At the core of our method are local features represented in the form of triplanes [9]. This representation is constructed as three perpendicular cross-planes, where each plane models the 3D surroundings from one perspective and by merging them, a thorough description of the 3D scene can be achieved. NeO 360's image-conditional tri-planar representation efficiently encodes information from image-level features while offering a compact queryable representation for any world point. We use these features combined with the residual local image-level features to optimize multiple unbounded 3D scenes from a large collection of images. NeO 360's 3D scene representation can build a strong prior for complete 3D scenes, which enables efficient 360° novel view synthesis for outdoor scenes from just a few posed RGB images.

To enable building a strong prior representation of unbounded outdoor scenes and given the scarcity of available multi-view data to train methods like NeRF, we also present

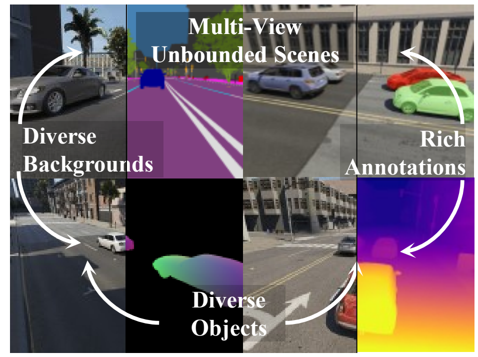

Figure 3: Samples from our large scale **NeRDS 360**: "**Ne**RF for **R**econstruction, **D**ecomposition and Scene **S**ynthesis of 360° outdoor scenes" dataset comprising 75 unbounded scenes with full multi-view annotations and diverse scenes.

a new large scale 360° unbounded dataset (Figure 3) comprising of more than 70 scenes across 3 different maps. We demonstrate our approach's effectiveness on this challenging multi-view unbounded dataset in both few-shot novel-view synthesis and prior-based sampling tasks. In addition to learning a strong 3D representation for complete scenes, our method also allows for inference-time pruning of rays using 3D ground truth bounding boxes, thus enabling compositional scene synthesis from a few input views. In summary, we make the following contributions:

- **A generalizable NeRF architecture for outdoor scenes** based on tri-planar representation to extend the NeRF formulation for effective few-shot novel view synthesis of 360° unbounded environments.
- **A large scale synthetic 360° dataset, called NeRDS 360, for 3D urban scene understanding** comprising multiple objects, capturing high-fidelity outdoor scenes with dense camera viewpoint annotations.
- Our proposed approach significantly outperforms all baselines for few-shot novel view synthesis on the NeRDS 360 dataset, showing 1.89 PNSR and 0.11 SSIM absolute improvement number for the 3-view novel-view synthesis task.

2. Related Works

Neural Implicit Representations use neural networks to map euclidean or temporal coordinates to target scene properties [57]. These methods have been successfully used to represent 3D shapes with signed distances [45, 15, 22, 20, 71] or occupancies [33, 56]. While earlier methods require ground truth 3D supervision, recent advances in differentiable neural rendering have enabled self-supervised learning of the target signal from only image supervision [52, 35], with Neural Radiance Fields (NeRFs) achieving impressive results, particularly for novel view synthe-

Datasets	NeRF[35]	T&T[26]	NeRF360[4]	BMVS[67]	MP3D[10]	DTU[2]	**Ours**
# Scenes	8	15	9	502	90	80	75
# Images	0.8k	2.8k	2k	101k	190k	3.9k	15k
Multi-Object	✗	✗	✗	✓	✓	✗	✓
Outdoor scenes	✗	✓	✓	✓	✗	✗	✓
Compositionality	✗	✗	✗	✗	✗	✓	✓
360 Camera	✗	✓	✓	✗	✓	✗	✓
All GT Annotations	✓	✗	✗	✗	✗	✗	✓

Table 1: **Comparison of NeRDS 360 with prior novel view synthesis datasets** comparing # scenes, scale, whether scenes contain multi-objects, are outdoor i.e. urban setting, supports compositionality, provides 360 cameras (as opposed to front-facing), and all ground truth information is available (*e.g.*, depth, 3D bounding boxes, instance masks)

sis. NeRF extensions focus on reducing aliasing effects via multiscale representations [3], modeling unbounded scenes [32], disentangled object-background representations and blending [42], compositional generative models [39] and improving reconstruction and depth estimation accuracy via multi-view consistent features [53, 54, 18].

Generalizable and Feature-conditioned representations: Neural implicit representations operate in the overfitting context, aiming to accurately encode a single scene/object [33, 35, 56, 37]. Other approaches use global conditioning (a single global latent vector) to learn object latent spaces [45, 23, 36, 49, 22], struggling to capture detailed high-frequency information. While usually limited to modeling single scenes, Pixel-NeRF [69], Scene Representation Transformers [50], and GRF [59] use local features as conditioners for generalizable neural fields. Recently, grounplanar [51] and tri-planar representations gain popularity for efficiently representing scenes using hybrid implicit-explicit representation. EG3D [9] and GAUDI [5] utilize tri-planar representations in their generative models, employing adversarial training. Notably, GAUDI utilizes GAN inversion for conditional synthesis. Our approach, NeO 360, does not rely on expensive GAN inversion and directly outputs density and RGB for new views in a feedforward manner.

Novel View Synthesis Datasets: Existing datasets for novel view synthesis fall under the following major categories: 1. Synthetic scenes which hemispherical 360-degree views around an object of interest offering dense camera overlap for fine-grained reconstruction (these include SRN rendering [12], RTMV [58] and Google Scanned Objects [14]), 2. Forward-facing scenes that move the camera in the vicinity of an object without providing full 360-degree coverage (these include LLFF [34], DTU [2], Blender MVS [67]) and 3. 360-degree real-scenes which provide full surrounding coverage (these include Tanks and Templates [26], MipNeRF360 dataset [3] and CO3D [47]). These datasets mostly evaluate on indoor scenes and provide little or no compositionality (i.e. multi-objects, 3D bounding boxes) for training or evaluation. While MipNeRF360 [3] and Tanks and Tempalte [26] provide 360 coverage, the number of scenes in these datasets is small; hence it is difficult evaluating the performance of generalizable NeRF methods at scale. Due to these challenges, we collect a large-scale outdoor dataset offering similar camera distributions as NeRF [35] for 360° outdoor scenes. Our dataset, described in detail in (Section 3), offers dense viewpoint annotations for outdoor scenes and is significantly larger than existing outdoor datasets for novel-view synthesis. This allows building effective priors for large-scale scenes which can lead to improved generalizable performance on new scenes with very limited views, as we show in Section 5.

3. NeRDS 360 Multi-View Dataset for Outdoor Scenes

Due to the challenge of obtaining accurate ground-truth 3D and 2D information (such as denser viewpoint annotations, 3D bounding boxes, and semantic and instance maps), only a handful of outdoor scenes have been available for training and testing. Specifically, previous formulations [44, 16, 28, 48] have focused on reconstructions using existing outdoor scene datasets [17, 31, 7] offering panoramic-views from the camera mounted on an ego-vehicle. These datasets [8, 55] offer little overlap between adjacent camera views [65], a characteristic known to be useful for training NeRFs and multi-view reconstruction methods. Moreover, the optimization of object-based neural radiance models for these scenes becomes more challenging as the ego car is moving fast and the object of interest is observed in just a few views (usually less than 5).

Dataset: To address these challenges, we present a large-scale dataset for 3D urban scene understanding. Compared to existing datasets, as demonstrated in Table 1, our dataset consists of 75 outdoor urban scenes with diverse backgrounds, encompassing over 15,000 images. These scenes offer 360° hemispherical views, capturing diverse foreground objects illuminated under various lighting conditions. Additionally, our dataset (as shown in Fig. 4, 5) encompasses scenes that are not limited to forward driving views, addressing the limitations of previous datasets such as limited overlap and coverage between camera views [17, 31]). The closest pre-existing dataset for generalizable evaluation is DTU [2] (80 scenes) which comprises mostly indoor objects and does not provide multiple foreground objects or background scenes.

We use the Parallel Domain [1] synthetic data generation to render high-fidelity 360° scenes. We select 3 different maps i.e. *SF 6thAndMission*, *SF GrantAndCalifornia* and *SF VanNessAveAndTurkSt* and sample 75 different scenes across all 3 maps as our backgrounds (All 75 scenes across 3 maps are significantly different road scenes from each other, captured at different viewpoints in the city). We select 20 different cars in 50 different textures for training and randomly sample from 1 to 4 cars to render in

Figure 4: **Proposed multi-view dataset** RGB renderings for 360° novel-view synthesis of outdoor scenes.

Figure 5: Camera distribution for 1, 3 and 5 source views (shown in red) and evaluation views (shown in green)

a scene. We refer to this dataset as **NeRDS 360**: **Ne**RF for **R**econstruction, **D**ecomposition and Scene **S**ynthesis of 360° outdoor scenes. In total, we generate 15k renderings (Fig. 4) by sampling 200 cameras in a hemispherical dome at a fixed distance from the center of cars. We held out 5 scenes with 4 different cars and different backgrounds for testing, comprising 100 cameras distributed uniformly sampled in the upper hemisphere, different from the camera distributions used for training. We use the diverse validation camera distribution to test our approach's ability to generalize to unseen viewpoints as well as unseen scenes during training. As shown in Figure 5 and supplementary material, our dataset and the corresponding task is extremely challenging due to occlusions, diversity of backgrounds, and rendered objects with various lightning and shadows. Our task entails reconstructing 360° hemispherical views of complete scenes using a handful of observations i.e. 1 to 5 as shown by red cameras in Fig. 5 whereas evaluating using all 100 hemispherical views, shown as green cameras in Fig. 5; hence our task requires strong priors for novel view synthesis of outdoor scenes.

4. Method

Given RGB images of a few views of a novel scene, NeO 360 infers a 3D scene representation capable of performing novel view synthesis and rendering 360° scenes. To achieve this goal, we employ a hybrid local and global feature representation comprised of a triplanar representation that can be queried for any world point. Formally, as shown in Fig. 1, given a few input images, $I = [I_1 ... I_n]$ of a complex scene, where $n = 1$ to 5, and their corresponding camera poses, $\gamma = [\gamma_1 ... \gamma_n]$ where $\gamma = [R|T]$, NeO 360 infers the density and radiance fields for both near and far backgrounds (similar to NeRF++ [73]) with the major difference of using hybrid local and global features for conditioning the radiance field decoders instead of just positions and viewing directions, as employed in the classical NeRF formulation [35, 73]. We describe our 3D scene representation in Section 4.1, introduce deep residual local features in Section 4.2, describe how we decode radiance fields conditioned on hybrid local and global features in Section 4.3, and discuss performing inference-time scene editing and composition in Section 4.4.

4.1. Image-Conditional Triplanar Representation

Preliminaries: NeRF is an implicit 3D scene representation that learns a neural network $f(\mathbf{x}, \theta) \rightarrow (\mathbf{c}, \sigma)$. This end-to-end differentiable function f outputs color c_i and density σ_i for every query 3D position x_i and the viewing direction θ_i as input. For each point evaluation, a 4 channel $(\mathbf{c}, \sigma)$ value is output, which is then alpha composited (Eq. 1 below) to render an image using volume rendering.

$$\mathbf{c} = \sum_i w_i \mathbf{c}_i, \quad \mathbf{acc} = \sum_i w_i \tag{1}$$

$$w_i = \alpha_i \prod_{j<i} (1 - \alpha_j), \quad \alpha_i = 1 - \exp\left(-\sigma_i \left\| \boldsymbol{x}_i - \boldsymbol{x}_{i+1} \right\|\right)$$

Method: Although producing high-fidelity scene synthesis, NeRF [35] is limited in its ability to generalize to novel scenes. In order to effectively use scene-priors and learn from a large collection of unbounded 360° scenes, we propose an image-conditional triplanar representation, as shown in Fig. 6. This representation is capable of modeling 3D scenes with full expressivity at scale without omitting any of its dimensions (as in 2D or BEV-based representation) and avoiding cubic complexity (as in voxel-based representations). Our triplanar representation comprises of three axis-aligned orthogonal planes $S = [\mathbf{S}_{xy}, \mathbf{S}_{xz}, \mathbf{S}_{yz}], \in \mathbb{R}^{3 \times C \times D \times D}$ where $D \times D$ is the spatial resolution of each plane with feature C.

To construct feature triplanes from input image, we first extract low-resolution spatial feature representations by using an ImageNet [13] pre-trained ConvNet backbone $\mathbf{E}$ which transforms input image $I \in \mathbb{R}^{H_i \times W_i \times 3}$ to 2D feature map $F_I \in \mathbb{R}^{H_i/2 \times W_i/2 \times C}$. Similar to prior works involving volumetric reconstruction [54, 38, 25], the obtained local features are projected backwards along every ray to the 3D feature volume (V_F) using camera pose γ_i and intrinsic K_i. While volumetric reconstruction methods [54, 38], traditionally use the generated volume solely for indoor geometry reconstruction through TSDF, we show that it can also be employed in a computationally efficient way to estimate the entire scene's appearance and enable accurate neu-

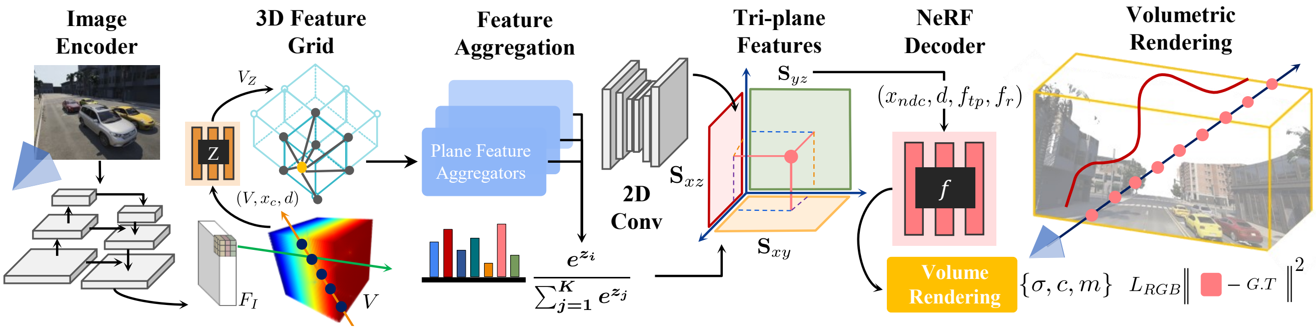

Figure 6: **Method:** Our method effectively uses local features to infer an image-conditional triplanar representation for both backgrounds and foregrounds. These triplanar features are obtained after orthogonally projecting positions (x) into each plane and bilinearly interpolating feature vectors. Dedicated NeRF decoder MLPs are used to regress density and color each for foreground and background.

ral rendering. Since all features along a camera ray are identical in the grid, we further learn depth of individual features by an additional MLP, $V_Z = Z(V_F, x_c, d)$ which takes as input concatenated features in the grid, positions of grid in the camera frame (x_c) and directions from the positions of grid in the world frame x_w to the camera frame and outputs depth-encoded features V_Z. Next, we obtain triplane features using learnt weights (w_i) over individual volumetric feature dimensions:

$$\mathbf{S}_{xy} = \sum_i w_{xy_i} \mathbf{V}_{Z_i}, \quad w_{xy} = A_{xy}(V_{Z_i}, x_{w_z}) \tag{2}$$

$$\mathbf{S}_{xz} = \sum_i w_{xz_i} \mathbf{V}_{Z_i}, \quad w_{xz} = A_{xz}(V_{Z_j}, x_{w_y}) \tag{3}$$

$$\mathbf{S}_{yz} = \sum_i w_{yz_i} \mathbf{V}_{Z_i}, \quad w_{yz} = A_{yz}(V_{Z_j}, x_{w_x}) \tag{4}$$

where A_{xy}, A_{xz} and A_{yz} denote feature aggregation MLPs and w_{xy}, w_{xz} and w_{yz} are softmax scores obtained after summing over the z, y and x dimensions respectively. One motivation to project features into respective planes is to avoid the computationally cubic complexity of 3D CNNs as in [11, 53] and at the same time be more expressive than BEV or 2D feature representations [29, 30, 69] which are computationally more efficient than voxel-based representations but omitting z-axis hurts their expressiveness. We instead rely on 2D convolutions to transform the built image-conditional triplanes into a new G-channel output, where $G = C/4$, while upsampling the spatial dimension of planes from $D \times D$ to image feature space (i.e. $H/2 \times W/2$). The learnt convolutions act as inpainting networks to fill in missing features. As shown in Fig. 6, our triplanar representation acts as a global feature representation, as intuitively a complex scene can be better represented when examined from various perspectives. This is because each may offer complementary information that can help understand the scene more effectively.

4.2. Deep Residual Local Features:

As noted by [11, 63] and inspired by [19], for the following radiance field decoding stage, we also use the features f_r as a residual connection into the rendering MLP. We obtain f_r from F_I by projecting the world point x into source view using its camera parameters γ_i, K_i and extracting features at the projected pixel locations through bilinear interpolation similar to [69]. Note that both local and global feature extraction pathways share the same weights θ_E and encoder $\mathbf{E}$. We find that for complex urban unbounded scenes, using just local features similar to [69] leads to ineffective performance for occlusions and faraway 360° views, as we show quantitatively and qualitatively in Section 5. Using only global features, on the other hand, leads to hallucinations, as shown in our ablations 5. Our method combines both local and global feature representations effectively, resulting in a more accurate 360° view synthesis from as minimal as a single view of an unbounded scene.

4.3. Decoding Radiance Fields:

The radiance field decoder D is tasked with predicting color c and density σ for any arbitrary 3D location x and viewing direction d from triplanes S and residual features f_r. We use a modular implementation of rendering MLPs, as proposed by [73] with the major difference of using our local and global features for conditioning instead of just using positions and viewing diretions as an input to the MLPs. The MLP is denoted as:

$$\sigma, c = D(x, d, f_{tp}, f_r) \tag{5}$$

where we obtain f_{tp} by orthogonally projecting point x into each plane in S and performing bi-linear sampling. We concatenate the three bi-linearly sampled vectors into $f_{tp} = [\mathbf{S}_{xy}(i,j), \mathbf{S}_{xz}(j,k), \mathbf{S}_{yz}(i,k)]$. Note that similar to [69], we establish our coordinate system using the view space of the input image, and then indicate the positions and camera rays within this particular coordinate system.

By utilizing the view space, our method can successfully standardize the scales of scenes from various data sources, thereby enhancing its ability to generalize well. Although our method gives reasonable results from single-view observation (Section 5), NeO 360 can seamlessly integrate multi-view observations by pooling along the view dimension in the rendering MLPs. We refer to our supplementary material for a detailed architecture diagram and description.

Near and Far Decoding MLPs: Similar to NeRF++ [73], we define two rendering MLPs for decoding color and density information as follows:

$$D(.) = \begin{cases} D_{fg}(.) & \text{if } \sum_{i=1}^{n}(x_i^2 + y_i^2 + z_i^2) < 1 \\ D_{bg}(.) & \text{if } \sum_{i=1}^{n}(x_i^2 + y_i^2 + z_i^2) > 1 \end{cases} \tag{6}$$

where we define a coordinate remapping function (M) similar to the original NeRF++ formulation [73] to contract the 3D points that lie outside the unit sphere where M maps points (x, y, z) outside the unit sphere to the new 4D coordinates$(x', y', z', 1/r)$, where (x', y', z') represents the unit vector in the direction of (x, y, z) and r denotes the inverse radius along this dimension. This formulation helps further objects get less resolution in the rendering MLPs. For querying our tri-planar representation, we use the un-contracted coordinates (x, y, z) in the actual world coordinates since our representation is planes instead of spheres. For rendering, we use the respective contracted coordinates (x', y', z') for conditioning the MLPs.

Optimizing radiance fields for few-shot novel-view synthesis: Given local and global features constructed from source views, we decode color c_p^i and density σ_p^i for backgrounds using dedicated near and far background MLPs $D_{near}(.)$ and $D_{far}(.)$ (Eq. 5) after volumetrically rendering and compositing the near and far backgrounds and enforcing the loss as follows:

$$L = \|c_p - \tilde{c}_t\|_2^2 + \lambda_{reg} L_{reg} + \lambda_{LPIPS} L_{LPIPS} \tag{7}$$

where $\tilde{c}_t$ is the sampled pixel locations from the target image and c^p is the composited color obtained from the rendering output of near and far MLPs as $c_i^p = c_i^{nb} + \prod_{j<i}\left(1 - \alpha_j^{nb}\right) c_i^b$. We also encourage the weights of near and far background MLPs to be sparse for efficient rendering by enforcing an additional distortion regularization loss [3] and further use L_{LPIPS} loss to encourage perceptual similarity b/w patches of rendered color, c^p and ground color $\tilde{c}_t$, where we only enforce it after 30 training epochs to improve background modeling (c.f. see supplementary).

4.4. Scene Editing and Decomposition:

Given 3D bounding boxes obtained from a detector, we can obtain the individual object and background radiance fields for each object by simply sampling rays inside the 3D bounding boxes of the objects and bilinearly interpolating the features at those specific (x, y, z) locations in our tri-planar feature grid ($\mathbf{S}$), making it straightforward to edit out and re-render individual objects. As illustrated in Fig. 7, we perform accurate object re-rendering by considering the features inside the 3D bounding boxes of objects to render the foreground MLP. In essence, we divide the combined editable scene rendering formulation as rendering objects, near backgrounds, and far backgrounds. For far backgrounds, we retrieve the scene color c_i^b and density σ_i^b which is unchanged from the original rendering formulation. For near backgrounds, we obtain color c_i^{nb} and density σ_i^{nb} after pruning rays inside the 3D bounding boxes of objects (i.e. setting σ_i^{nb} to a negative high value, -1×10^{-5} before volumetrically rendering). For objects, we only consider rays inside the bounding boxes of each object and sampling inside foreground MLP to retrive c_i^o and density σ_i^o. We aggregate the individual opacities and colors along the ray to render composited color using the following equation:

$$\mathbf{c} = \sum_i w_i^b \mathbf{c}_i^b + \sum_i w_i^{nb} \mathbf{c}_i^{nb} + \sum_i w_i^o \mathbf{c}_i^o \tag{8}$$

5. Experiments

We evaluate our proposed method against various state-of-the-art baselines, focusing on few-shot novel view synthesis including **a.** Conditional prior-based sampling and **b.** Novel scene rendering tasks. We compare full scenes on the following baselines: 1) **NeRF** [35]: Vanilla NeRF formulation which overfits to a scene given posed RGB images 2) **PixelNeRF** [69] A generalizable NeRF variant which utilizes local image features for few-shot novel view synthesis 3) **MVSNeRF** [11]: Extends NeRF for few-view synthesis using local features obtained by building a cost-volume from source images and 4) **NeO 360**: Our proposed architecture which combines local and global features for generalizable scene representation learning.

Metrics: We use standard PSNR, SSIM, and LPIPS metrics to evaluate novel-view synthesis and L1 and RMSE to measure depth reconstruction quality.

Comparison with strong baselines for novel-view synthesis: We aim to answer the following key questions: **1.** Does our generalizable tri-planar representation perform better than other generalizable NeRF variants given access to prior data and a few views for optimization on novel scenes? **2.** Do scene priors help with zero-shot generalization? and **3.** Does scaling the data help our network generalize better? We summarize the results in Table 2 and note that NeO 360 achieves superior results compared to state-of-the-art generalizable NeRF variants i.e. PixelNeRF [69] and MVSNeRF [11] in both zero-shot testing and finetuning given a limited number of source views. Spefically NeO

Method	# Views	Single Map (Prior Sampling)				Single Map (Novel Scenes)				Multi-Map (Novel Scenes)			
		Scenes			**Objects**	**Scenes**			**Objects**	**Scenes**			**Objects**
		PSNR ↑	**SSIM** ↑	**LPIPS** ↓	**PSNR** ↑	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**PSNR** ↑	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**PSNR** ↑
▲ NeRF [35]	1	12.41	0.09	0.71	10.55	12.82	0.13	0.69	10.45	12.72	0.12	0.69	10.39
	3	14.14	0.25	0.59	11.67	13.76	0.18	0.62	11.81	13.82	0.19	0.61	11.85
	5	15.37	0.38	0.50	12.89	16.16	0.39	0.49	14.83	16.14	0.38	0.48	14.73
▲ MVS-NeRF [11]	1	14.40	0.40	0.65	12.42	13.87	0.33	0.65	11.08	13.90	0.31	0.65	11.10
	3	13.93	0.34	0.63	11.33	14.50	0.40	0.64	12.79	14.40	0.38	0.63	12.70
	5	14.78	0.39	0.62	12.21	15.43	0.41	0.61	14.13	15.40	0.42	0.62	14.10
▲ Pixel-NeRF [69]	1	15.89	0.44	0.64	13.57	14.93	0.40	0.65	12.93	15.01	0.47	0.65	12.65
	3	17.15	0.50	0.62	14.47	17.46	0.48	0.63	15.70	16.20	0.52	0.64	13.00
	5	17.50	0.51	0.62	14.75	17.80	0.49	0.62	15.92	16.91	0.52	0.62	14.22
▲ Pixel-NeRF$_{ft}$ [69]	1	-	-	-	-	15.78	0.43	0.66	14.07	14.65	0.42	0.66	11.81
	3	-	-	-	-	17.90	0.51	0.59	17.12	16.69	0.54	0.62	13.49
	5	-	-	-	-	19.26	0.55	0.57	18.54	17.22	0.55	0.61	15.21
▲ **NeO 360** (Ours)	1	16.91	0.51	0.56	14.11	17.60	0.56	0.51	15.80	16.30	0.52	0.57	13.04
	3	18.94	0.58	0.48	16.66	19.35	0.59	0.50	17.60	18.59	0.61	0.52	15.93
	5	19.64	0.62	0.47	17.34	20.10	0.62	0.48	18.20	19.27	0.64	0.49	16.60
▲ **NeO 360$_{ft}$** (Ours)	1	-	-	-	-	17.93	0.58	0.49	15.95	16.42	0.55	0.54	13.80
	3	-	-	-	-	19.56	0.61	0.46	18.30	18.94	0.63	0.49	16.81
	5	-	-	-	-	20.56	0.64	0.45	18.62	**19.59**	**0.67**	**0.46**	**17.70**

Table 2: **Quantitative novel view synthesis results**: Conditional prior-based sampling and novel-scene rendering. Pink ▲ denotes zero-shot evaluation whereas Orange ▲ denotes finetuning only the triplanar network i.e. freezing the encoder **E** with learning rate 5×10^{-6} from 1,3 or 5 source views. The orange triangle for PixelNeRF denotes that we finetune their encoder network similar to how we finetune our triplanes. Underline shows best results when trained and evaluated within the same single map (easier setting), **Bold** denotes best results with challenging evaluation setting where the evaluation dataset is sampled across 3 different maps with diverse illumination.

Method	**Backgrounds**		**Objects**
	PSNR↑	**SSIM**↑	**PSNR**↑
NeRF [34] (No Priors)	16.16	0.34	15.42
Ours (No Pretraining)	16.40	0.39	15.70
Ours	**20.48**	**0.67**	**19.03**

Table 3: Effect of scene priors on 3-view novel view synthesis on NeRDS 360 dataset (Single Map)

Method	**PSNR**↑	**SSIM**↑	**LPIPS**↓
with Colors	16.29	0.44	0.62
w/o Feature Grid	17.50	0.47	0.59
w/o Near/Far	19.02	0.57	0.52
Ours (global + local)	**19.35**	**0.59**	**0.50**

Table 4: Effect of different design choices in our architecture for 3-view novel view synthesis on NeRDS 360 dataset (Single Map)

360 achieves a PSNR of 19.35, SSIM of 0.59, and LPIPS of 0.50 for complete scenes and 17.60 PSNR for objects with zero-shot evaluation, hence demonstrating an absolute improvement of 1.89 PSNR, 0.11 SSIM, 0.13 LPIPS for complete scenes and 1.90 PSNR for novel objects against the best-performing baseline for a single-map scenario (i.e. all methods trained on 25 scenes and evaluated on a single novel scene within the same map). Our approach also outperforms the best baselines on the challenging multi-map dataset with zero-shot evaluation, where 5 novel scenes were held out with difficult illuminations and shadows. NeO 360 achieves an absolute PSNR improvement of 2.39 on complete scenes and 2.93 on objects, showing that NeO 360 learns better priors for unbounded scenes. The table also shows both NeO 360 and PixelNeRF perform better than the original NeRF since NeRF has not seen any scene level prior and is optimized per scene from only a few images.

Ablation Analysis: We further show the effect of our design choices in Table 4. We show that directly using colors from source views results in a much worse performance of our model due to over-reliance on source view pixels which hurts generalization to out-of-distribution camera views. Additionally, we ablate the feature grid as well as the near/far MLPs. The results confirm that performance degrades without the use of the 3D feature grid i.e. global features (row 2) which is one of our major contributions. The near/far decomposition has a less significant but still positive effect. Overall, our model with combined local and global features in the form of triplanes performs the best among all variants.

Effect of Scene Priors: Table 2 shows NeO 360's ability to overfit to a large number of scenes, hence achieving better PSNR than PixelNeRF on sampling from prior distributions for novel trajectories. To further validate the effect of scene priors, we omit scene priors from our approach's training and the results are summarized in Table 3. The results further confirm that scene priors actually help our network, resulting in an absolute PSNR improvement of 4.08 comparing the network learned on 25 scenes with the network which has seen no prior scenes during training and is only overfit on 3 unseen views from scratch for a novel scene. The results show that the architecture design of our network allows us to learn from a large collection of scenes

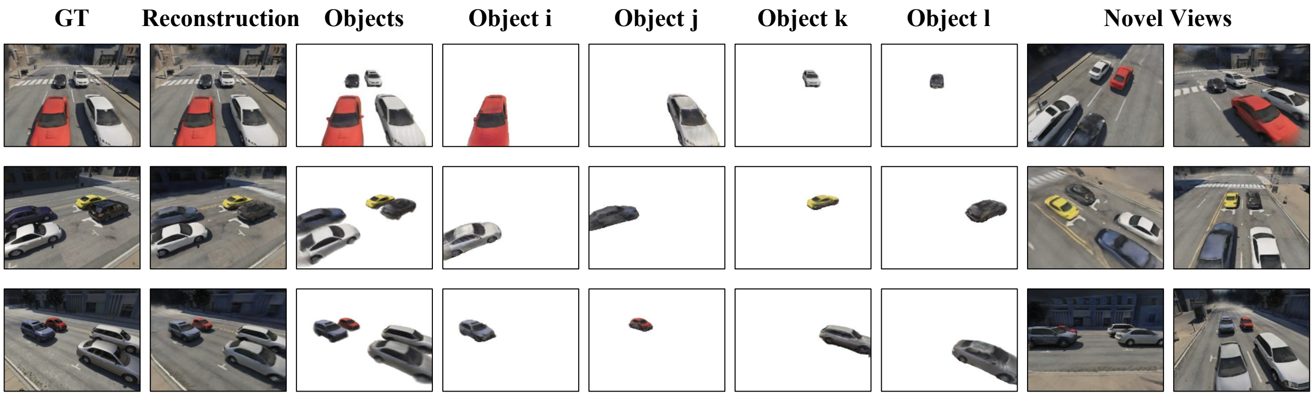

Figure 7: **Scene decomposition qualitative results** showing 3-view scene decomposed individual objects along with novel views on the NeRDS 360 dataset. Our approach performs accurate decomposition by sampling inside the 3D bounding boxes of the objects; hence giving full control over object editability from very few input views.

Comparison	PSNR ↑	SSIM ↑
mipNeRF360 [4]	13.25	0.31
SRT [50]	14.61	0.40
EG3D*/GAUDI* [9] [5]	12.84	0.30
NeO 360 (Ours)	**19.35**	**0.59**

Table 5: Quantitative comparison of 3-view view synthesis on NeRDS 360

Object Depth	L1 ↓	RMSE ↓
PixelNeRF [70]	0.83	1.07
NeO 360 (no ft.)	0.59	0.74
NeO 360 (Ours)	**0.20**	**0.61**

Table 6: Eval views depth prediction from 3 source views on NeRDS 360

while extending the learned prior to novel scenes with effective zero-shot generalizability from a few views.

Additional baseline comparisons: We include additional comparisons with novel-view synthesis baselines (mipNeRF360 [4], EG3D [9] and SRT [50]) in Tab. 5. One could clearly observe that naively using triplanes (Tab. 5 row 3, *denotes that we take the triplanar representation without generative losses or training) or local features as in PixelNeRF (Tab. 4) hurt the performance. Our method relies on effective combination of local and global features which serves as a strong baseline for the challenging task of 360° view synthesis of outdoor scenes. We also include depth reconstruction metrics (Tab. 6) and show our techniques' superior results compared to Pixel-NeRF.

6. Qualitative Results:

Comparison with generalizable NeRF baselines: As seen in Figure 8, our method excels in novel-view synthesis from 3 source views, outperforming strong generalizable NeRF baselines. Vanilla NeRF struggles due to overfitting on these 3 views. MVSNeRF, although generalizable, is limited to nearby views as stated in the original paper, and thus struggles with distant views in this more challenging task whereas PixelNeRF's renderings also produce artifacts for far backgrounds.

Detailed comparison with PixelNeRF: Figure 9 presents our method's novel-view synthesis results compared with PixelNeRF. The red and blue boxes focus on close-ups. The visual comparison emphasizes our method's ability to produce crisper and clearer object and background renderings, while PixelNeRF generates blurrier out-

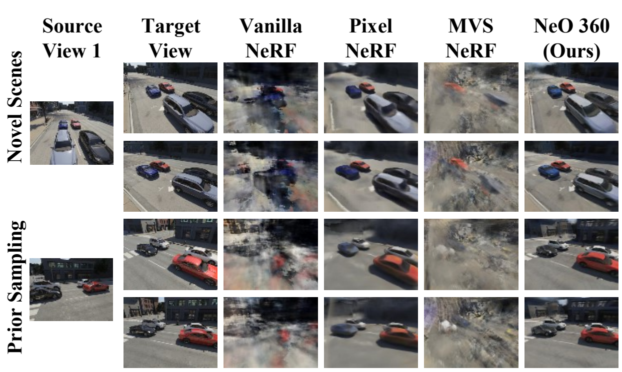

Figure 8: **Qualitative 3-view view synthesis results**: Comparisons with baselines.

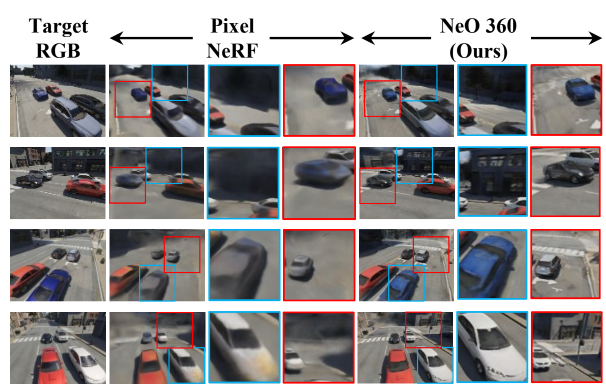

Figure 9: **Qualitative 3-view view synthesis results**: Close-up comparison with PixelNeRF [69].

Figure 10: **NeO 360's zero-shot qualitative results:** We show 360° predictions for 3-view and 5-view novel view synthesis. Note that although our network has some shape artifacts for 3-view novel-view synthesis, these are effectively resolved by adding a few more sparse views, showing our network's ability to effectively use learned priors for sparse novel-view synthesis in a zero-shot manner. We show 10 predicted samples with indices 11, 20, 32, 38, 43, 48, 65, 76, 84, and 98 rendered from a circular trajectory generated at a consistent radius around the scene.

puts with noticeable artifacts in both foreground and background.

Scene Decomposition: We further show our network's scene decomposition performance in Figure 7. The figure demonstrates precise object recovery from the near background MLP output through sampling within each object's GT 3D bounding box (as emphasized in Section 4.4). This formulation allows us to easily re-render objects thanks to our feature-based representation which can be queried individually for objects and backgrounds. Note that we do not enforce any objectness prior during training to get this behavior, it is purely learned from multi-view image-based rendering.

360° qualitative results: We further show our network's predicted few-view 360° novel view synthesis output in a zero-shot manner on unseen scenes and objects, not observed during training. As shown by Fig. 10, our method performs plausible novel-view synthesis of complete scenes including far-away backgrounds and objects from very few sparse views of an outdoor scene; hence demonstrating NeO 360's ability to use learned priors effectively. We also show that 3-view synthesis introduces some artifacts in parts of the scene where there is no overlap between source views i.e. where the scene is entirely unobserved. Fig. 10 shows that by adding a few sparse sets of views in those areas (i.e. 5-view case), those artifacts can be effectively removed and a smooth scene representation could be obtained. This shows our network's ability to interpolate smoothly across given source views and also complete the scene in an effective manner.

7. Conclusion

In this paper, we proposed NeO 360, a generalizable extension to the NeRF approach for unbounded 360° scenes. Our method relies on image-conditional tri-planar representations for few-shot novel view synthesis. In order to build strong priors for unbounded scenes, we propose a large-scale dataset, NERDS 360 to study view synthesis, reconstruction and decomposition in a 360-degree setting. Our method performs significantly better than other generalizable NeRF variants and achieves higher performance when tested on novel scenes. For future work, we will explore how the proposed method can be used to build priors that rely less on labelled data, such as 3D bounding boxes during inference and instead rely on motion cues for effective scene decomposition without labelled data. A second avenue of future work consists of sim2real extensions of this work to alleviate the data and annotations requirement in the real world by using only labelled data in simulation.

References

[1] Parallel domain. https://paralleldomain.com/, May 2021. 3

[2] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016. 3

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 3, 6

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2, 3, 8

[5] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*, 2022. 3, 8

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *CVPR*, 2021. 1

[7] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv*, 2020. 3

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv*, 2019. 3

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2, 3, 8

[10] 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. *International Conference on 3D Vision (3DV)*, 2017. 3

[11] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 5, 6, 7

[12] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[14] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 3

[15] Yueqi Duan, Haidong Zhu, He Wang, Li Yi, Ram Nevatia, and Leonidas J Guibas. Curriculum deepsdf. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 51–67. Springer, 2020. 2

[16] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *International Conference on 3D Vision (3DV)*, 2022. 1, 3

[17] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

[18] Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rare Ambru, Greg Shakhnarovich, Matthew R Walter, and Adrien Gaidon. Depth field networks for generalizable multi-view scene representation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 245–262. Springer, 2022. 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[20] Nick Heppert, Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Rares Andrei Ambrus, Jeannette Bohg, Abhinav Valada, and Thomas Kollar. Carto: Category and joint agnostic reconstruction of articulated objects. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21201–21210, 2023. 2

[21] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 1

[22] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi object shape appearance and pose optimization. 2022. 1, 2, 3

[23] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2, 3

[24] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video, 2022. 2

[25] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017. 4

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

[27] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, pages 5742–5752. PMLR, 2021. 2

[28] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 2, 3

[29] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 5

[30] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022. 5

[31] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[32] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 3

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2, 3

[34] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 3, 7

[35] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 6, 7

[36] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022. 2, 3

[37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[38] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020. 4

[39] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[40] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 2

[41] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. In *Robotics: Science and Systems*, 2022. 1

[42] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 1, 3

[43] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2

[44] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2856–2865, June 2021. 3

[45] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2, 3

[46] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo-Martin Brualla. Deformable neural radiance fields. *arXiv*, 2020. 1

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

[48] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022. 3

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

[50] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. *CVPR*, 2022. 2, 3, 8

[51] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T Freeman, Fredo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. Seeing 3d objects in a single image via self-supervised static-dynamic disentanglement. *arXiv preprint arXiv:2207.11232*, 2022. 3

[52] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 2019. 2

[53] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 3, 5

[54] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 3, 4

[55] 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. 3

[56] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. 2021. 2, 3

[57] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, volume 41, pages 703–735. Wiley Online Library, 2022. 1, 2

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

[59] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3

[60] Itsuki Ueda, Yoshihiro Fukuhara, Hirokatsu Kataoka, Hiroaki Aizawa, Hidehiko Shishido, and Itaru Kitahara. Neural density-distance fields. *arXiv preprint arXiv:2207.14455*, 2022. 1, 2

[61] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022. 2

[62] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2

[63] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2, 5

[64] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond.(2021), 2021. 1

[65] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. S-nerf: Neural radiance fields for street views. *arXiv preprint arXiv:2303.00749*, 2023. 3

[66] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, October 2021. 2

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

[68] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[69] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 3, 5, 6, 7, 8

[70] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 8

[71] Sergey Zakharov, Rares Ambrus, Katherine Liu, and Adrien Gaidon. Road: Learning an implicit recursive octree auto-decoder to efficiently encode 3d shapes. *arXiv preprint arXiv:2212.06193*, 2022. 2

[72] Sergey Zakharov, Rares Andrei Ambrus, Vitor Campagnolo Guizilini, Dennis Park, Wadim Kehl, Fredo Durand, Joshua B Tenenbaum, Vincent Sitzmann, Jiajun Wu, and Adrien Gaidon. Single-shot scene reconstruction. In *5th Annual Conference on Robot Learning*, 2021. 2

[73] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 2, 4, 5, 6