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

Muhammad Zubair Irshad\textsuperscript{1,2} Sergey Zakharov\textsuperscript{2} Katherine Liu\textsuperscript{2} Vitor Guizilini\textsuperscript{2} Thomas Kollar\textsuperscript{2}
Adrien Gaidon\textsuperscript{2} Zsolt Kira\textsuperscript{*1} Rares Ambrus\textsuperscript{*2}

\textsuperscript{1}Georgia Institute of Technology \hspace{1em} \textsuperscript{2}Toyota Research Institute

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

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 triplane 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-
DeepSDF [45]
Backgrounds
Diverse
Unbounded Scenes
x
reconstructing appearance and shapes. The
taxonomy of implicit representation methods
Figure 2: Taxonomy of implicit representation methods
constraining 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 represen-
tation as well as the computational expense of per-scene
optimization from many views, we propose to infer the rep-
resentation 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 presented 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 represen-
tation 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 un-
bounded outdoor scenes and given the scarcity of available
multi-view data to train methods like NeRF, we also present

<table>
<thead>
<tr>
<th>Multi-View Unbounded Scenes</th>
<th>DeepSDF [45]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diverse Backgrounds</td>
<td>ShAPo [22]</td>
</tr>
<tr>
<td>Multi-View Unbounded Scenes</td>
<td>CodeNeRF [23]</td>
</tr>
<tr>
<td>Diverse Objects</td>
<td>AutoRF [36]</td>
</tr>
<tr>
<td>Rich Annotations</td>
<td>CARTO [20]</td>
</tr>
</tbody>
</table>

Figure 3: Samples from our large scale NeRDS 360: "NeRF for Reconstruction, Decomposition and Scene Synthesis of 360° outdoor scenes" dataset comprising 75 unbounded scenes with full multi-view annotations and diverse scenes.

• 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 prop-
erties [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 re-
quire ground truth 3D supervision, recent advances in dif-
ferentiable neural rendering have enabled self-supervised
learning of the target signal from only image supervi-
sion [52, 35], with Neural Radiance Fields (NeRFs) achieving
impressive results, particularly for novel view synthe-
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, groupplanar [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 Tempaltes [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 Tempaltes [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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Scenes</td>
<td>8</td>
<td>15</td>
<td>9</td>
<td>162</td>
<td>80</td>
<td>40</td>
<td>75</td>
</tr>
<tr>
<td># Images</td>
<td>0.8k</td>
<td>2.8k</td>
<td>2k</td>
<td>101k</td>
<td>190k</td>
<td>3.9k</td>
<td>15k</td>
</tr>
<tr>
<td>Multi-Object</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outdoor scenes</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compositionality</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>360 Camera</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All GT Annotations</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td>✓  ✓  ✓  ✓  ✓  ✓  ✓  ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).
in Fig. 1, given a few input images, $I$, can be queried for any world point. Formally, as shown this goal, we employ a hybrid local and global feature novel view synthesis and rendering 360 infers a 3D scene representation capable of performing 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, $γ = [γ_1...γ_n]$ where $γ = [R|T]$, 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. Method

4.1. Image-Conditional Triplanar Representation

**Preliminaries:** NeRF is an implicit 3D scene representation that learns a neural network $f(x, θ) → (c, σ)$. 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 $(c, σ)$ value is output, which is then alpha compositized (Eq. 1 below) to render an image using volume rendering.

$$\begin{align*}
  c &= \sum_i w_i c_i, \\
  \text{acc} &= \sum_i w_i \\
  w_i &= α_i \prod_{j < i} (1 - α_j), \quad α_i = 1 - \exp(-σ_i \| x_i - x_{i+1} \|)
\end{align*}$$

**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 $S = [S_{xy}, S_{xz}, S_{yz}]$, $∈ \mathbb{R}^{3×C×D×D}$ where $D × D$ 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 $E$ which transforms input image $I ∈ \mathbb{R}^{H_i × W_i × 3}$ to 2D feature map $F_I ∈ \mathbb{R}^{H_f / 2 × W_f / 2 × C}$. 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-
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.

**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 \(\gamma_i, K_i\) 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 \(\theta_E\) and encoder \(E\). 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 \(\sigma\) 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 directions as an input to the MLPs. The MLP is denoted as:

\[
\sigma, c = D(x, d, f_{ip}, f_r)
\]

where we obtain \(f_{ip}\) by orthogonally projecting point \(x\) into each plane in \(S\) and performing bi-linear sampling. We concatenate the three bi-linearly sampled vectors into \(f_{ip} = [S_{xy}(i,j), S_{xz}(j,k), S_{yz}(i,k)]\). 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.

---

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.

<table>
<thead>
<tr>
<th><strong>Method:</strong></th>
<th><strong>3D Feature Grid</strong></th>
<th><strong>Feature Aggregation</strong></th>
<th><strong>Tri-plane Features</strong></th>
<th><strong>NeRF Decoder</strong></th>
<th><strong>Volumetric Rendering</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Image Encoder</strong></td>
<td><strong>Voxelization</strong></td>
<td><strong>Planar Feature Aggregators</strong></td>
<td><strong>NeRF Decoder</strong></td>
<td><strong>View Synthesis</strong></td>
<td><strong>Volume Rendering</strong></td>
</tr>
</tbody>
</table>
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 \(c_i\) and density \(\sigma_i\) 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:

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

where \(\tilde{c}_i\) 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 \(c_i = c_i^{nb} + \prod_{j \in S} (1 - \sigma_j^{nb}) \sigma_i^b\). 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 \(c_i\), 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 \(S\), 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 \(\sigma_i^b\) after pruning rays inside the 3D bounding boxes of objects (i.e. setting \(\sigma_i^{nb}\) to a negative high value, \(-1 \times 10^{-5}\) before volumetrically rendering). For objects, we only consider rays inside the bounding boxes of each object and sampling inside foreground MLP to retrieve \(c_i^b\) and density \(\sigma_i^b\). We aggregate the individual opacities and colors along the ray to render composited color using the following equation:

\[
c = \sum_i w_i^b c_i^b + \sum_i w_i^{nb} c_i^{nb} + \sum_i w_i^o c_i^o \tag{8}
\]

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. Specifically NeO
shows best results when trained and evaluated within the same single map (easier setting),
0.64
0.47
0.45
19.59 0.67 0.46 17.70
18.62
20.10 0.62 0.48 18.20 19.27 0.64 0.49 16.60
prior and is optimized per scene from only a few images.
The original NeRF since NeRF has not seen any scene level
shows both NeO 360 and PixelNeRF perform better than
learns better priors for unbounded scenes. The table also
complete scenes and 2.93 on objects, showing that NeO 360
were held out with difficult illuminations and shadows. NeO
outperforms the best baselines on the challenging multi-
map dataset with zero-shot evaluation, where 5 novel scenes
all methods trained on 25 scenes and evaluated on a sin-
Table 3: Effect of scene priors on 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 pix-

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

Table 5: Quantitative comparison of 3-view view synthesis 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 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.

**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


[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


[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

