IBRNet: Learning Multi-View Image-Based Rendering

Qianqian Wang\textsuperscript{1,2}, Zhicheng Wang\textsuperscript{1} Kyle Genova\textsuperscript{1,3}, Pratul Srinivasan\textsuperscript{1} Howard Zhou\textsuperscript{1} Jonathan T. Barron\textsuperscript{1} Ricardo Martin-Brualla\textsuperscript{1} Noah Snavely\textsuperscript{1,2} Thomas Funkhouser\textsuperscript{1,3}

\textsuperscript{1}Google Research \textsuperscript{2}Cornell Tech, Cornell University \textsuperscript{3}Princeton University

Abstract

We present a method that synthesizes novel views of complex scenes by interpolating a sparse set of nearby views. The core of our method is a network architecture that includes a multilayer perceptron and a ray transformer that estimates radiance and volume density at continuous 5D locations (3D spatial locations and 2D viewing directions), drawing appearance information on the fly from multiple source views. By drawing on source views at render time, our method hearkens back to classic work on image-based rendering (IBR), and allows us to render high-resolution imagery. Unlike neural scene representation work that optimizes per-scene functions for rendering, we learn a generic view interpolation function that generalizes to novel scenes. We render images using classic volume rendering, which is fully differentiable and allows us to train using only multi-view posed images as supervision. Experiments show that our method outperforms recent novel view synthesis methods that also seek to generalize to novel scenes. Further, if fine-tuned on each scene, our method is competitive with state-of-the-art single-scene neural rendering methods.

1. Introduction

Given a set of posed images of a scene, the goal of novel view synthesis is to produce photo-realistic images of the same scene at novel viewpoints. Early work on novel view synthesis focused on image-based rendering (IBR). Starting from the pioneering work on view interpolation of Chen and Williams [6], and proceeding through light field rendering [3, 15, 29], view-dependent texturing [8], and more modern learning-based methods [17], IBR methods generally operate by warping, resampling, and/or blending source views to target viewpoints. Such methods can allow for high-resolution rendering, but generally require either very dense input views or explicit proxy geometry, which is difficult to estimate with high quality leading to artifacts in rendering.

More recently, one of the most promising research directions for novel view synthesis is neural scene representations, which represent scenes as the weights of neural networks. This research area has seen significant progress through the use of Neural Radiance Fields (NeRF) [39]. NeRF shows that multi-layer perceptrons (MLPs) combined with positional encoding can be used to represent the continuous 5D radiance field of a scene, enabling photo-realistic novel view synthesis on complex real-world scenes. NeRF’s use of continuous scene modeling via MLPs, as opposed to explicit discretized volumes [51] or multi-plane images [12, 65] allows for more compact representations and scalability to larger viewing volumes.

Although neural scene representations like NeRF can represent scenes faithfully and compactly, they typically require a lengthy optimization process for each new scene before they can synthesize any novel views of that scene, which limits the value of these methods for many real-world applications.

In this work, we leverage ideas from both IBR and NeRF into a new learning-based method that generates a continuous scene radiance field on-the-fly from multiple source views for rendering novel views. We learn a general view interpolation function that simultaneously performs density/occlusion/visibility reasoning and color blending while rendering a ray. This enables our system to operate without any scene-specific optimization or precomputed proxy geometry.

At the core of our method is a lightweight MLP network that we call IBRNet, which aggregates information from source views along a given ray to compute its final color. For sampled 3D locations along the ray, the network first fetches latent 2D features derived from the source views that encode spatial context. IBRNet then aggregates the 2D features from source views for each sample to produce a density feature that captures information about whether that feature seems to be on a surface. A ray transformer module then computes a scalar density value for each sample by considering these density features along the entire ray, enabling visibility reasoning across larger spatial scales. Separately, a

\url{https://ibrnet.github.io/}
color blending module uses the 2D features and view direction vectors from source views to compute a view-dependent color for each sample, which is represented as a weighted combination of the projected colors of the source views. A final color value is then computed for each ray using volume rendering.

Our approach is fully differentiable and can therefore be trained end-to-end using multi-view images. Our experiments show that when trained on large amounts of data, our method can render high-resolution photo-realistic novel views for unseen scenes that contain complex geometry and materials, and our quantitative evaluation shows that it improves upon state-of-the-art novel view synthesis methods designed to generalize in a single shot to new test scenes. Moreover, for any particular scene, we can fine-tune IBRNet to improve the quality of synthesized novel views to match the performance of state-of-the-art neural scene representation methods like NeRF [39]. In summary, our contributions are:

- a new learning-based multi-view image-based rendering approach that outperforms existing one-shot view synthesis methods on novel scenes,
- a new model architecture called IBRNet that enables the continuous prediction of colors and densities in space from multiple views,
- a per-scene fine-tuning procedure that achieves comparable performance to state-of-the-art novel view synthesis methods designed only for single-scene inference.

2. Related work

Image based rendering. Early work on IBR introduced the idea of synthesizing novel views from a set of reference images by a weighted blending of reference pixels [9, 15, 29]. Blending weights were computed based on ray-space proximity [29] or approximate proxy geometry [3, 9, 19]. In more recent work, researchers have proposed improved methods for computing proxy geometry [5, 18], optical flow correction [4, 10, 11], and soft blending [42, 47]. For example, Hedman et al. [17] use two types of multi-view stereo [22, 49] to first produce a view-dependent mesh surface and then a CNN to compute blending weights. Others synthesize a radiance field directly on a mesh surface [8, 21, 57] or point cloud [1, 57, 43]. While these methods can handle sparser views than other approaches and achieve promising results in some cases, they are fundamentally limited by the performance of 3D reconstruction algorithms [22, 49]. They have difficulty in low-textured or reflective regions, where stereo reconstruction tends to fail, and they cannot handle partially translucent surfaces. In contrast, our method learns continuous volume densities in an end-to-end manner that is optimized for synthesis quality, leading to better performance in challenging scenarios.

Volumetric Representations. Another line of work uses discrete volumetric representations to achieve photo-realistic rendering. Recent methods leverage convolutional neural networks (CNNs) to predict volumetric representations stored in voxel grids [20, 26, 27, 42, 58] or multi-plane images (MPIs) [12, 13, 35, 38, 54, 65]. At test time, novel views can be rendered from these representations via alpha compositing [44]. Trained end-to-end on large datasets, these methods often learn to compensate for the discretization artifacts of low-resolution voxel grids and can generalize reasonably well to different scenes. While they achieve high-quality results for novel view synthesis, they must explicitly process and store large numbers of samples resulting in extensive memory and computational overhead, limiting the resolution of their outputs. In contrast, our method allows for querying color and opacity at continuous 3D locations and 2D viewing directions without storing a full scene representation, and can easily scale to rendering high-resolution images. Our method can also handle larger viewing volumes than MPI-based methods.

Neural scene representations. A recent promising research direction is the use of neural networks for representing the shape and appearance of scenes. Earlier work [2, 14, 23, 35, 41, 53, 61] showed that MLPs can be used as implicit shape representations, where the weights of the MLPs map continuous spatial coordinates to signed distance or occupancy values. With advances in differentiable rendering methods [7, 24, 30, 33, 34, 40], many methods [32, 35, 40, 48, 50, 51, 52, 56, 62] showed the ability to learn both scene geometry and appearance from multi-view observations (though only for simple geometry and diffuse materials). The more recent NeRF [39] achieves very impressive results for novel view synthesis by optimizing a 5D neural radiance field for a scene, suggesting the advantages of continuous representations over discrete ones like voxel grids or meshes. However, NeRF must be optimized for each new scene, taking hours or days to converge. Concurrent work [58, 63] tries to address this issue and generalize NeRF (with a focus on very sparse input views). However, their use of absolute locations as direct network inputs restricts their ability to generalize to arbitrary new scenes. In contrast, our method can generalize to new, real-world scenes with high quality.

3. Method

Given nearby source views, our method uses volume rendering to synthesize a target view at a novel camera pose. The core problem we try to solve is to obtain colors and densities in a continuous space by aggregating information present in the source views. Our system pipeline (Fig. 1) can be divided into three parts: 1) identifying a set of nearby source views as input and extracting dense features from
While our network (described in the next section) can handle an arbitrary number of neighboring views, given limited GPU memory we select a small number of source views as the “effective working set” for rendering a novel view. To obtain an effective working set, we identify spatially close candidate views, then select the subset of \( N \) views whose viewing directions are most similar to the target view.

Let \( \mathbf{I}_i \in [0, 1]^{H_i \times W_i \times 3} \) and \( \mathbf{P}_i \in \mathbb{R}^{3 \times 4} \) respectively denote the color image and camera projection matrix for the \( i \)-th source view. We use a shared U-Net based convolutional neural network to extract dense features \( \mathbf{F}_i \in \mathbb{R}^{H_i \times W_i \times d} \) from each image \( \mathbf{I}_i \). The set of tuples \( \{(\mathbf{I}_i, \mathbf{P}_i, \mathbf{F}_i)\}_{i=1}^N \) forms the input for rendering a target view.

### 3.2. RGB-σ prediction using IBRNet

Our method synthesizes images based on classic volume rendering: accumulating colors and densities in the 3D scene to render 2D images. We propose IBRNet (illustrated in Fig. 2) to predict colors and densities at continuous 5D locations by aggregating information from multiple source views and incorporating long-range context along the ray. Our proposed IBRNet is permutation-invariant and accepts a variable number of source views.

We first describe the input to IBRNet for a single query point location \( \mathbf{x} \in \mathbb{R}^3 \) on a ray \( \mathbf{r} \in \mathbb{R}^3 \) with unit-length viewing direction \( \mathbf{d} \in \mathbb{R}^3 \). We project \( \mathbf{x} \) into all source views using their camera parameters, and extract colors and features at the projected pixel locations through bilinear interpolation. Let \( \{\mathbf{C}_i\}_{i=1}^N \in [0, 1]^3 \) and \( \{\mathbf{f}_i\}_{i=1}^N \in \mathbb{R}^d \) respectively denote these extracted colors and image features for point \( \mathbf{x} \). We also take into account the viewing directions for \( \mathbf{x} \) in all source views, denoted as \( \{\mathbf{d}_i\}_{i=1}^N \).

#### 3.2.1 Volume density prediction

Our density prediction at point \((\mathbf{x}, \mathbf{d})\) involves two steps. First, we aggregate the multi-view features at \((\mathbf{x}, \mathbf{d})\) to obtain a density feature. Then, our proposed ray transformer takes in density features for all samples on a ray and incorporates long-range contextual information to predict the density for each sample including \((\mathbf{x}, \mathbf{d})\).

**Multi-view feature aggregation.** We observe that 3D points on surfaces are more likely to have consistent local appearances in multiple views than 3D points in free space. Therefore, an effective way to infer density would be to check the consistency among the features \( \{\mathbf{f}_i\}_{i=1}^N \) for a given point, and a possible implementation would be a PointNet-like architecture that takes in multi-view features and
uses variance as the global pooling operator. However, such a simple global consistency check might erroneously fail for a true surface in the presence of occlusions. Instead, we construct a small network to allow for better occlusion reasoning.

Specifically, we first compute a per-element mean \( \mu \in \mathbb{R}^d \) and variance \( \nu \in \mathbb{R}^d \) from features \( \{ f_i \}_{i=1}^N \) to capture global information, and concatenate each \( f_i \) with \( \mu \) and \( \nu \). Each concatenated feature is fed into a small shared MLP to integrate both local and global information, resulting in a multi-view aware feature \( f'_i \) and a weight vector \( w_i \in [0, 1] \). We pool these new features \( \{ f'_i \}_{i=1}^N \) by computing their weighted average and variance using weights \( \{ w_i \}_{i=1}^N \), which is mapped to a density feature \( f_\sigma \in \mathbb{R}^{d_\sigma} \) using an MLP. Compared to a direct average or max-pooling in a PointNet [45], we find that our weighted pooling improves the network's ability to handle occlusions.

**Ray transformer.** After obtaining a density feature \( f_\sigma \), one could directly turn it into a single density \( \sigma \) with another MLP. However, we find such an approach fails to predict accurate densities for new scenes with complex geometry. We ascribe this to the fact that looking at features for a single point in isolation is inadequate for accurately predicting its density, and more contextual information is needed—similar to how plane-sweep stereo methods consider matching scores along a whole ray before determining the depth of a particular pixel.

We therefore introduce a new ray transformer module to enable samples on a ray to attend to each other before predicting their densities. The ray transformer consists of the two core components of a classic Transformer [60]: positional encoding and self-attention. Given \( M \) samples along the ray, our ray transformer treats samples from near to far as a sequence, and applies positional encoding and multi-head self-attention to the sequence of density features \( \{ f_i(x_1) \cdots f_i(x_M) \} \). The final density value \( \sigma \) for each sample is then predicted from its attended feature. The ray transformer module allows for a variable number of inputs and only introduces only a small increase in parameters and computational overhead, while dramatically improving the quality of the predicted densities and the final synthesized image, as shown in our ablation studies.

**Improving temporal visual consistency.** Our method considers only nearby source views as the working set when synthesizing a target view. Therefore, when generating videos along smooth camera paths, our method is potentially subject to temporarily inconsistent density predictions and flickering artifacts due to abrupt changes in the working set as the camera moves.

To alleviate this issue, we adopt the pooling technique introduced by Sun et al. [55]. We replace the mean \( \mu \) and variance \( \nu \) of \( \{ f_i \}_{i=1}^N \) with a weighted mean \( \mu_w \) and variance \( \nu_w \), weighted so as to reduce the influence of the furthest images in the working set. The weight function is defined as
follows:
\[
\hat{w}_i^f(\mathbf{d}, \mathbf{d}_i) = \max \left(0, e^{s(\mathbf{d} \cdot \mathbf{d}_i - 1)} - \min_{j=1 \ldots N} e^{s(\mathbf{d} \cdot \mathbf{d}_j - 1)} \right),
\]
\[
w_i^{f}(\mathbf{d}, \mathbf{d}_i) = \frac{\hat{w}_i^f(\mathbf{d}, \mathbf{d}_i)}{\sum_j \hat{w}_j^f(\mathbf{d}, \mathbf{d}_j)},
\]
where \(s\) is a learnable parameter. This technique improves our method in two ways: 1) it smooths the change in global features between adjacent frames, and 2) it produces more reasonable global features by up-weighting closer views in the working set and down-weighting more distant views. We observe this technique empirically improves synthesis stability and quality.

### 3.2.2 Color prediction

We obtain color at a 5D point by predicting blending weights for the image colors \(\{C_i\}_{i=1}^{N}\) in the source views that correspond to that 5D point. Unlike NeRF [39], which uses absolute viewing directions, we consider viewing direction relative to that of the source views, i.e., the difference between \(\mathbf{d}\) and \(\mathbf{d}_i\). A smaller difference between \(\mathbf{d}\) and \(\mathbf{d}_i\) typically implies a larger chance of the color at target view resembling the corresponding color at view \(i\), and vice versa.

To predict the blending weight for each source color \(C_i\), we concatenate the feature \(f_i^c\) with \(\Delta \mathbf{d}_i = \mathbf{d} - \mathbf{d}_i\) and input each concatenated feature into a small network to yield a blending weight \(w_i^c\). The final color for this 5D location is blended via a soft-argmax operator \(c = \sum_i^N [C_i \exp(w_i^c)] / \sum_j^N \exp(w_j^c)]\). An alternative to predicting blending weights would be to directly regress \(c\). However, we found that direct regression of \(c\) led to decreased performance.

### 3.3. Rendering and training

The approach introduced in the previous section allows us to compute the color and density value at continuous 5D location. To render the color of a ray \(\mathbf{r}\) passing through the scene, we first query the colors and densities of \(M\) samples on the ray, and then perform classic volume rendering [25] with the integral approximated using numerical quadrature [39]:

\[
\hat{C}(\mathbf{r}) = \sum_{k=1}^{M} T_k (1 - \exp(-\sigma_k \delta_k)) \mathbf{c}_k,
\]

where
\[
T_k = \exp\left(-\sum_{j=1}^{k-1} \sigma_j \delta_j\right).
\]

Hierarchical volume rendering. One benefit of having continuous RGB-\(\sigma\) predictions is that it allows more adaptive and efficient sampling in space. Following NeRF [39], we perform hierarchical volume sampling and simultaneously optimize two networks, a coarse IBRNet and a fine IBRNet, with identical network architecture. At the coarse scale, we sample a set of \(M_c\) locations at equidistant disparity (inverse depth) which results in equal intervals between adjacent point projections in pixel space. Given the prediction of the coarse network, we then conduct a more informed sampling of points along each ray, where samples are more likely to be located at relevant regions for rendering. We sample an additional \(M_f\) locations and use all \(M_c + M_f\) locations to render the fine results as in NeRF [39]. In our network, we set both \(M_c\) and \(M_f\) to 64.

Training objective. We render the color of each ray using both the coarse and fine set of samples, and minimize the mean squared error between the rendered colors and ground-truth pixel colors for training:

\[
\mathcal{L} = \sum_{\mathbf{r} \in \mathcal{R}} \left(\left\| \hat{C}_c(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 + \left\| \hat{C}_f(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2\right)
\]

where \(\mathcal{R}\) is the set of rays in each training batch. This allows us to train the feature extraction network as well as the coarse and fine IBRNet simultaneously.

Fine-tuning. Trained on large amounts of data, our method is able to generalize well to novel scenes. Our method can also be fine-tuned per scene using the same objective at training time (Eq. 4) but only on a specific scene, thereby improving synthesis performance on that scene.

### 3.4. Implementation details

Source and target views sampling. Given multiple views of a scene, we construct a training pair of source and target view by first randomly selecting a target view, and then sampling \(N\) nearby views as source views. To select source views, we first identify a pool of \(n \cdot N\) nearby views (\(n\) is sampled uniformly at random from \([1, 3]\)), and then randomly sample \(N\) views from the pool. This sampling strategy simulates various view densities during training and therefore helps the network generalize across view densities. During training, we sample \(N\) uniformly at random from \([8, 12]\).

Network details. We implement the image feature extraction network using a U-Net like architecture with ResNet34 [16] truncated after layer3 as the encoder, and two additional up-sampling layers with convolutions and skip-connections as the decoder. We decode two sets of feature maps in the final decoding layer, to be used as input to the coarse and fine IBRNet respectively. Both coarse and fine feature maps have \(d = 32\) dimensions, and are \(1/4 \times \) the original image size. For IBRNet, the dimension of the density feature \(f_o\) is 16 and we use 4 heads for the self-attention
module in ray transformer. Please refer to the supplementary material for more network details.

**Training details** We train both the feature extraction network and the IBRNet end-to-end on datasets of multi-view posed images using Adam [28]. The base learning rates for feature extraction network and IBRNet are $10^{-3}$ and $5 \times 10^{-4}$, respectively, which decay exponentially along with the optimization. During finetuning, we optimize both our 2D feature extractor and IBRNet itself using a smaller base learning rate ($5 \times 10^{-4}$ and $2 \times 10^{-4}$). We train on eight V100 GPUs with a batch size of 3,200 to 9,600 rays depending on image resolution and the number of source views.

### 4. Experiments

We evaluate our method in two ways: a) directly evaluating our pre-trained model on all test scenes without any fine-tuning (Ours). Note that we train only one model and evaluate it on all test scenes. And b) fine-tuning our pre-trained model on each test scene before evaluation (Ours$_{fin}$), similar to NeRF.

#### 4.1. Experimental Settings

**Training datasets.** Our training data consists of both synthetic data and real data. For synthetic data, we generate object-centric renderings of the 1,030 models in Google Scanned Objects [46] at a variety of view density rates. For real data, we use RealEstate10K [65], the Spaces dataset [12], and 95 real scenes from handheld cellphone captures (36 from LLFF [38] and 59 from ourselves). RealEstate10K is a large dataset of indoor scenes with camera poses derived from about 80K video clips, where the cameras lie on 1D camera trajectories. The Spaces dataset contains 100 scenes, where each is captured by a 16-camera rig at 3 to 10 different rig positions. Each of the cellphone captured scenes has 20-60 images captured by forward-facing cameras distributed roughly on 2D grids [38]. We use COLMAP [49] to estimate camera poses, intrinsic parameters, and scene bounds for our captures. Our training data includes various camera setups and scene types, which allows our method to generalize well to unseen scenarios.

**Evaluation datasets.** We use both synthetic rendering of objects and real images of complex scenes for evaluation, as in NeRF [39]. The synthetic data consists of four Lambertian objects with simple geometry from the DeepVoxels [51] dataset, and eight objects with complicated geometry and realistic materials from the NeRF synthetic dataset. Each object in DeepVoxels has 479 training views and 1,000 test views at $512 \times 512$ resolution sampled on the upper hemisphere. Objects in the NeRF synthetic dataset have 100 training views and 200 test views at $800 \times 800$ resolution sampled either on the upper hemisphere or the full sphere. The real dataset contains 8 complex real-world scenes captured with roughly forward-facing images from NeRF. Each scene is captured with 20 to 62 images at 1008 x 756 resolution. 1/8th of the images are held out for testing.

**Baselines.** We compare our method against the current best view synthesis methods from two classes: **MPI-based methods** and **neural rendering methods**. For MPI-based methods, we compare Ours with a state-of-the-art method LLFF [38] which uses a 3D convolutional neural network to predict a discretized grid for each input image and blends nearby MPIs into the novel viewpoint. Both Ours and LLFF are designed to generalize to novel scenes. We compare Ours$_{fin}$ with neural rendering methods: Neural Volumes (NV) [35], Scene Representation Networks (SRN) [52], and NeRF [39]. These methods train separate networks for each scene or each class of scenes, and either do not generalize at all, or only generalize within object categories. When evaluating against these neural scene representation methods we fine-tune our model per-scene, as do the baselines.

<table>
<thead>
<tr>
<th>Method</th>
<th>Settings</th>
<th>Diffuse Synthetic 360°</th>
<th>Realistic Synthetic 360°</th>
<th>Real Forward-Facing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PSNR $\uparrow$</td>
<td>SSIM $\uparrow$</td>
<td>LPIPS $\downarrow$</td>
</tr>
<tr>
<td>LLFF [38]</td>
<td>No per-scene optimization</td>
<td>34.38 0.985 0.048</td>
<td>24.88 0.911 0.114</td>
<td>24.13 0.798 0.212</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>37.17 0.990 0.017</td>
<td>25.49 0.916 0.100</td>
<td>25.13 0.817 0.205</td>
</tr>
<tr>
<td>SRN [52]</td>
<td>Per-scene optimization</td>
<td>33.20 0.963 0.073</td>
<td>22.26 0.846 0.170</td>
<td>22.84 0.668 0.378</td>
</tr>
<tr>
<td>NV [35]</td>
<td></td>
<td>29.62 0.929 0.099</td>
<td>26.05 0.893 0.160</td>
<td>- - -</td>
</tr>
<tr>
<td>NeRF [39]</td>
<td></td>
<td>40.15 0.991 0.023</td>
<td>31.01 0.947 0.081</td>
<td>26.50 0.811 0.250</td>
</tr>
<tr>
<td>Ours$_{fin}$</td>
<td></td>
<td>42.93 0.997 0.009</td>
<td>28.14 0.942 0.072</td>
<td>26.73 0.851 0.175</td>
</tr>
</tbody>
</table>

Table 1: **Quantitative comparison on datasets of synthetic and real images.** Our evaluation metrics are PSNR/SSIM (higher is better) and LPIPS [64] (lower is better). Both Ours and LLFF [38] are trained on large amounts of training data and then evaluated on all test scenes without any per-scene tuning. Ours consistently outperforms LLFF [38] on all datasets. We also compare our method with neural rendering methods (SRN [52], NV [35], and NeRF [39]) that train a separate network for each scene. To compete fairly with these methods, we also fine-tune our pre-trained model on each scene (Ours$_{fin}$). After finetuning, Ours$_{fin}$ is competitive with the state-of-the-art method NeRF [39].
Figure 3: **Qualitative comparison on Real-Forward-Facing data.** Our method can more accurately recover fine details in both geometry and appearance, and produce images that are perceptually more similar to ground-truth than other methods. LLFF [38] has difficulty recovering clean and accurate boundary (ghosting artifacts in Orchid and repeated edges in Horns), and fails to capture thin structures (ribs in T-Rex) or partially occluded origins (leaves in Fern). NeRF [39]-rendered images exhibits unrealistic noise in Orchid. Meanwhile, the texture on the petals is missing and the region near the boundary of the petals are not well recovered. Our method is also slightly better than NeRF at fine structures in T-Rex and Fern, and can reconstruct more details of the reflection on the glass in Horns.

4.2. Results

To render each test view we sample 10 source views from the training set for all evaluation datasets. We evaluate our method and our baselines using PSNR, SSIM, and LPIPS [64]. Results can be seen in Tab. 1 and in Fig. 3.

Tab. 1 shows that our pretrained model generalizes well to novel scenes and consistently outperforms LLFF [38] on all test scenes. After finetuning, our model’s performance is competitive with state-of-the-art neural rendering methods like NeRF [39]: We outperform NeRF on both Diffuse Synthetic 360° and Real Forward-Facing [38] and only have lower PSNR and SSIM on Realistic Synthetic 360°. One reason that we underperform NeRF on Realistic Synthetic 360° is that this data has very sparse training views for scenes with very complex geometry. Therefore, the information contained in the limited set of local source views may be insufficient for synthesizing a novel view, in which case methods that optimize a global radiance field using all views like NeRF may be better suited.

On the Real Forward-Facing [38] data Ours ft achieves substantially better SSIM and LPIPS than NeRF [39], indicating that our synthesized images look more photo-realistic. A key component in NeRF is the use of a positional encoding [56] which helps generate high-frequency details. But positional encoding may also cause unwanted high-frequency artifacts in images (see Fig. 3Orchid), which reduces perceptual quality. In contrast, our model does not use positional encoding to regress colors but instead blends colors from source views, which biases our synthesized images to look more like natural images. Our method produces reasonable proxy geometry and can be used to produce temporarily consistent videos. Please refer to the supplementary material for more detail.

4.3. Ablations and analysis

Ablation studies. We conduct ablation studies (Tab. 2) to investigate the individual contribution of key aspects of our method, using our pre-trained models on Real Forward-
Facing [38] data. For “No ray transformer”, we delete ray transformer so that the density prediction for samples on each ray is independent and local. Removing this module causes the network to fail to predict accurate densities, leading to “black hole” artifacts and blurriness (see supplementary material for qualitative comparison). For “No view directions”, we remove the view direction input in our network. This includes removing the weighted pooling in Eq. 1 and removing the viewing direction as a network input for color prediction. This reduces our model’s ability to reproduce view-dependent effects such as specularities, but does not reduce performance as significantly as seen in NeRF because our model blends nearby input images (which themselves contain view-dependent effects).

**Sensitivity to source view density.** We investigate how our method degrades as the source views become sparser on Diffuse Synthetic 360° [51] data. Each scene in the original data provides 479 of training views randomly distributed on the upper hemisphere. We uniformly sub-sample the training views by factors of 2, 4, 6, 8, 10 to create varying view densities. We report the PSNR for both the pre-trained model Ours and the fine-tuned model Ours_f in Fig. 4. We see that our method degrades reasonably as the provided views become sparser.

**Computational overhead at inference.** Here we investigate the computation required by IBRNet in comparison to other methods. Our inference pipeline for rendering a target view can be divided into two stages: We first extract image features from all source views of interest, which needs only be done once as features can then be used repeatedly for rendering. Given these pre-computed features we use our proposed IBRNet to produce the colors and densities for samples along each camera ray. Because image features can be precomputed, the major computational cost at inference is incurred by IBRNet. Since our method interpolates novel views from a set of source views, the #FLOPs of IBRNet depends on the size of the local working set, i.e., the number of source views (#Src.Views). We report the size of IBRNet and #FLOPs required to render a single pixel under varying numbers of source views in Tab. 3.

We see that IBRNet requires less than 20% of the #FLOPs that NeRF [39] requires while achieving comparable performance. There are two possible explanations for why IBRNet is potentially more efficient than NeRF. First, NeRF uses a single MLP to encode the radiance field of the whole scene. To query any point in the scene, one must “touch” every parameter of the MLP. As the scene becomes more complex (e.g., city scale), NeRF needs more network capacity, which leads to larger #FLOPs for each query. In contrast, our method interpolates nearby views locally, and therefore the #FLOPs does not grow with the scale of the scene but only with the size of the local working set. This property also makes our approach more dynamic, enabling applications like streaming long scenes and adjusting computational costs on-the-fly to suit a device. Second, we encode the scene as posed images and features which is a more structured and interpretable representation than the MLP used in NeRF. A simple projection operation allows us to fetch the most relevant information from source views for estimating radiance at a point in space. The network must only learn how to combine that information, not how to synthesize it from scratch.

<table>
<thead>
<tr>
<th>Method</th>
<th>#Params</th>
<th>#Src.Views</th>
<th>#FLOPs</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeRF</td>
<td>1.19M</td>
<td>-</td>
<td>304M</td>
<td>26.50</td>
<td>0.811</td>
<td>0.250</td>
</tr>
<tr>
<td>Ours</td>
<td>0.04M</td>
<td>8</td>
<td>45M</td>
<td>26.56</td>
<td>0.847</td>
<td>0.176</td>
</tr>
<tr>
<td>Ours_f</td>
<td>0.04M</td>
<td>10</td>
<td>55M</td>
<td>26.73</td>
<td>0.851</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 3: Network size and computational cost at inference. The network size of our MLP is much smaller than NeRF [39]. For both NeRF [39] and Ours, we report the #FLOPS needed to render a single pixel. Both methods use hierarchical volume sampling. $M_r$, $M_t$ are set to 64, 128 respectively for NeRF, and 64, 64 for Ours. For our method, the number of #FLOPS scales roughly linearly with the number of source views (#Src.Views).

Table 2: Ablation study on Real Forward-Facing [38] data. We report the metrics for the pre-trained model of each ablation without per-scene fine-tuning.

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>No ray transformer</td>
<td>21.31</td>
<td>0.675</td>
<td>0.355</td>
</tr>
<tr>
<td>No view directions</td>
<td>24.20</td>
<td>0.796</td>
<td>0.243</td>
</tr>
<tr>
<td>Full model Ours</td>
<td>25.13</td>
<td>0.817</td>
<td>0.205</td>
</tr>
</tbody>
</table>

Figure 4: Sensitivity to source view density. We uniformly subsample the views provided in Diffuse Synthetic 360° [51] by {2, 4, 6, 8, 10} to create source view densities varying from 479 to 47 views on the upper hemisphere.
5. Conclusion

We have proposed a learning-based multi-view image-based rendering framework that synthesizes novel views of a scene by blending pixels from nearby images with weights and volume densities inferred by a network comprising an MLP and ray transformer. This approach combines the advantages of IBR and NeRF to produce state-of-the-art rendering quality on complex scenes without requiring pre-computed geometry (unlike many IBR methods), storing expensive discretized volumes (unlike neural voxel representations), or expensive training for each new scene (unlike NeRF).

References
