

im2nerf: Image to Neural Radiance Field in the Wild

Lu Mi^{1,2,†}, Abhijit Kundu¹, David Ross¹, Frank Dellaert^{1,4}, Noah Snavely^{1,3}, Alireza Fathi¹
¹Google Research ²MIT, ³Cornell Tech, Cornell University ⁴Georgia Tech

Abstract

We propose *im2nerf*, a learning framework that predicts a continuous neural object representation given a single input image in the wild, supervised by only segmentation output from off-the-shelf recognition methods. The standard approach to constructing neural radiance fields [27] takes advantage of multi-view consistency and requires many calibrated views of a scene, a requirement that cannot be satisfied when learning on large-scale image data in the wild. We take a step towards addressing this shortcoming by introducing a model that encodes the input image into a disentangled object representation that contains a code for object shape, a code for object appearance, and an estimated camera pose from which the object image is captured. Our model conditions a NeRF on the predicted object representation and uses volume rendering to generate images from novel views. We train the model end-to-end on a large collection of input images. As the model is only provided with single-view images, the problem is highly under-constrained. Therefore, in addition to using a reconstruction loss on the synthesized input view, we use an auxiliary adversarial loss on the novel rendered views. Furthermore, we leverage object symmetry and cycle camera pose consistency. We conduct extensive quantitative and qualitative experiments on the ShapeNet dataset [5] as well as qualitative experiments on Open Images dataset [19]. We show that in all cases, *im2nerf* achieves the state-of-the-art performance for novel view synthesis from a single-view unposed image in the wild.

1. Introduction

In recent years, the computer vision community has made great strides towards learning accurate representation of the 3D shape and appearance of objects and scenes [27, 14, 30, 3, 22]. Learning such representations can result in more realistic view synthesis and image generation [42, 33], as well as richer 3D scene understanding [51, 30]. To learn 3D representations, prior work often takes advantage of

multi-view consistency, given multiple images of the scene or object from different views with known camera poses. However, the assumption that the model can have access to multiple posed views of scenes either during training or testing does not hold for large-scale image data in the wild.

A few recent approaches aim at predicting 3D shape and appearance of objects from unposed single-view images [48, 10]. However, the 3D object representations used (e.g., meshes or voxels combined with implicit representations) result in reconstructions that lack detail. Furthermore, the problem is often constrained by encouraging object instances to look similar to an average category template which results in a low level of deformation in the final predictions.

In this work, we propose **im2nerf**, a learning framework that maps a given single input image to a predicted continuous neural object representation that through volume rendering synthesizes novel views of the object. Our approach is based on an auto-encoder model. The encoder decomposes each input image in terms of object geometry, object appearance and camera pose. The decoder is a NeRF-like [27, 14] model that can render the input image or new views of the object. It generates rays corresponding to pixels in the image based on a given camera pose, and uses a multi-layer perceptron (MLP) conditioned on the predicted shape and appearance representations to predict the volume density and view-dependent radiance for each 3D point.

As we lack 3D supervision or multiple views of the in-the-wild images, inspired by some recent approaches [48, 10, 45, 13], we leverage other sources of meta-supervision present in our knowledge of the 3D world. Our encoder-decoder approach encourages the object geometry and appearance to be consistent with the input image through differentiable rendering. Furthermore, we use an adversarial loss to encourage the rendering of novel views of the object to look realistic. As the novel views are rendered conditioned on the same representation as the input view, multi-view consistency is forced across all the rendered images. We furthermore leverage the fact that many object categories have in principle a symmetric 3D shape. Finally, we propose a cycle consistency loss that indirectly constrains the estimated camera poses to be geometrically meaningful.

†Work done during internship at Google.

Corresponding to lumi@mit.edu

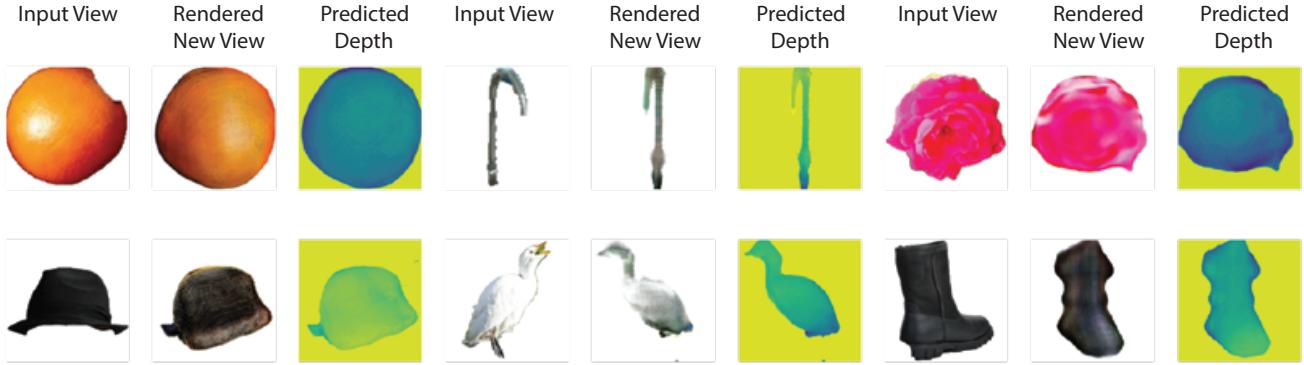


Figure 1. We propose im2nerf, a learning framework that predicts a continuous neural object representation given a single-view unposed input image in the wild.

ful. We feed the rendered novel view back to the encoder to re-estimate the camera pose.

To summarize, our contributions are as follows:

- We propose an unsupervised learning framework that predicts a continuous 3D neural representation given a single input image in the wild.
- We approach this highly under-constrained problem setting by leveraging auxiliary sources of meta-supervision including an adversarial loss on rendered novel views, using scene box, encouraging object symmetry and using cycle camera pose consistency.
- Our results achieve the state-of-the-art performance for novel view synthesis from a single-view unposed image in the wild.
- We further apply our method to a weakly supervised setting where we are provided with pose supervision only for a very small subset of the data. We show that our method is able to leverage the unlabeled data to significantly improve the model to achieve a comparable performance to a model trained with full supervision of pose annotations.

2. Related Work

Neural Representations. Using implicit functions to represent the 3D geometry and appearance of objects and scenes has gained popularity in recent years [32, 26, 28, 27, 14, 29, 3, 50, 42, 30, 22, 13, 2, 24]. These methods use an MLP as a function that maps the continuous spatial coordinates and view angles into signed distance, occupancy, color or density values. The advances in implicit representations together with the recent progress in differentiable rendering [16, 21, 15, 23, 31], have led to the emergence of NeRF [27], which captures viewpoint-varying image properties by storing volume density and view-dependent radiance fields from multi-view observations.

NeRF makes several assumptions: (a) the camera poses of input images and their intrinsics are known, (b) the model is optimized on a single static scene, and (c) many input images from multiple views are required to achieve the best results. Various extensions to NeRF have made progress on relaxing some of these assumptions. PixelNeRF [50], GRF [39] and IBRNet [42] can generalize to more than one scene by using an image encoder to condition the neural radiance field on image features. These methods still require multiple posed views of the scene during the training stage but can use a small number of views during the inference stage. iNeRF [49], NeRF-- [43], GNeRF [25] and BARF [22] aim at relaxing the known camera pose condition. However, these methods work best when the initial camera pose is close to the correct one. Henzler *et al.* [12] and Li *et al.* [20] learn a 3D object or scene representation from video. NeRF-VAE [17], GIRAFFE [30], Stelzner *et al.* [37] and DeVries *et al.* [6] learn to decompose the scene into objects or/and generate scenes with conditional radiance fields. CodeNeRF [14] uses an auto-decoder approach (similar to DeepSDF [32]) that jointly learns a code representation per object and the weights of the conditional NeRF multi-layer perceptron. All of these methods assume they are provided with multiple posed views of the objects or the scenes during the training stage, and do not show results on large-scale images in the wild.

Supervised Single-view 3D Reconstruction. Various data collection efforts have played a key role in early progress in 3D reconstruction from 2D images: synthetic datasets like ShapeNet [46] and SunCG [36], generation of synthetic data from real-world images [4, 52, 41], as well as other work that annotate real world images by fitting 3D CAD models to object silhouettes like Pascal3D [47] and Pix3D [38]. Some of the earlier methods leverage such data to formulate the problem assuming full 3D supervision [46, 5, 44, 34, 9, 8, 40, 11, 26]. However, requiring ground-truth 3D supervision restricts these methods to syn-

thetic or small datasets, since ground-truth 3D shapes are costly or impossible to acquire at large scale.

Unsupervised Single-view 3D Reconstruction. Progress in differentiable renderers [16, 21, 15, 23, 31] has allowed building end-to-end methods that lift an image into a 3D representation and render the 3D representation back to a novel view. DeepSDF [32] optimizes an implicit signed distance function that is used for representing the object in 3D by rendering it back to multiple images, requiring only 2D supervision. Goel *et al.* [10] predict a mesh-based 3D representations of objects from 2D images. However, the mesh-based representation often lacks the expressiveness needed to render high fidelity images. Wu *et al.* [45] explore using symmetry as a constraint to encourage more accurate 3D representations. PrGAN [7], HoloGAN [29], GRAF [35], pi-GAN [3] and GIRAFFE [30] use a generative model that is trained with an adversarial loss for learning 3D representation of objects from unposed images. These models are unconditional, meaning that they are able to generate an object or a scene given a random noise vector. However, they wouldn't necessarily be able to generate such representation conditioned on an input image.

The closest approach to ours is the work by Ye *et al.* [48], which proposes a method for learning 3D object representations in the wild. They use a combination of voxel and implicit representations to predict the 3D shape and appearance of the objects. However, the reconstructions usually lack details as a result of choices made in the 3D representation.

3. Method

An overview of our method is shown in Figure 2. Our goal is to learn a model that maps an input image segmented with an object mask to a 3D representation that can be rendered from novel views. Since the model should generalize to images in the wild, we do not assume any other sources of 2D or 3D supervision during training or testing. Our model consists of an encoder-decoder architecture. The encoder takes in the masked image of an object and predicts a latent shape and a latent appearance code, as well as the pose of the camera that captured the image (assuming the object is at the origin in its canonical frame and the camera intrinsics are known). The decoder takes in the latent codes and a camera pose (with known intrinsics) and renders a novel view image.

The key challenge in our problem setting is that we must optimize the parameters of the encoder and the conditional NeRF MLP in the decoder from a collection of unposed single-view images. Inspired by earlier work [48, 10, 3, 30] in this space, we propose to use meta-supervision that is present in our knowledge of the 3D world. We combine a differentiable rendering loss on the input view with an ad-

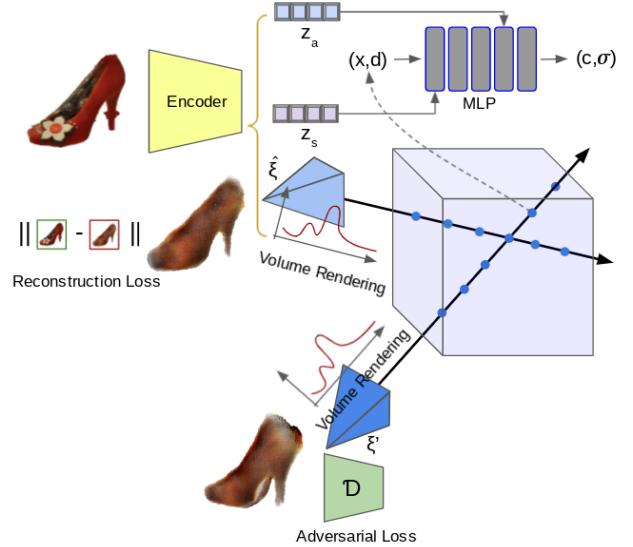


Figure 2. Overview of our method. Given an input image, the encoder predicts a shape z_s and an appearance code z_a and estimates the pose of the camera $\hat{\xi}$ that captures the input image. The decoder conditions a NeRF on the predicted shape and appearance representations and uses volume rendering to generate images from novel views. In addition to using a photometric reconstruction loss for input view, we apply an adversarial loss on rendered images from novel views. In addition, we further constrain the problem by using a scene box, cycle camera pose consistency and object symmetry (for symmetric object categories).

versarial loss on the novel views, together with the geometric consistency that is inherent in the conditional NeRF to train our model. Furthermore, we leverage other real-world knowledge such as scene box, object symmetry and cycle camera pose consistency loss to optimize our model.

3.1. Model

Our model follows a typical encoder-decoder structure, and the details of model components are introduced as follows,

Encoder. Our encoder consists of two feature extractor towers followed by prediction heads. The first tower predicts the shape and appearance latent codes for the object. The MLP in the decoder is conditioned on these codes to predict the color and density of the points in 3D from different view angles. The second tower predicts the pose of the camera that captures the input image. We assume the camera intrinsics are known and the world coordinate system is centred on the 3D object. The camera pose is described by a rotation and a translation matrix. To predict the rotation matrix, our network predicts 4 parameters: $\cos \theta, \sin \theta, \cos \phi, \sin \phi$ where θ is *azimuth* and ϕ is *elevation*. We then compute the rotation matrix as $\mathbf{R} = \mathbf{R}_\theta \mathbf{R}_\phi$. Our camera pose tower predicts the translation of the object in the cam-

era frame by predicting t_x , t_y and t_z . Since the camera is always looking at the object, t_z is always positive.

Decoder. The decoder receives three inputs—a camera pose, a shape code and an appearance code—and outputs a rendered image. It first generates a set of pixel rays based on the input camera pose. Points are sampled along the rays, and for each point, color and density values are estimated conditioned on the shape and appearance codes.

Conditioning on Shape and Appearance Codes. Inspired by the recent progress on generative NeRF models [3, 14, 17], we use a multi-layer perceptron that maps scene coordinates \mathbf{x} and view directions \mathbf{d} to their corresponding density $\hat{\sigma}$ and RGB color values $\hat{\mathbf{c}}$ conditioned on the shape and appearance codes

$$(\hat{\mathbf{c}}, \hat{\sigma}) = F(\gamma(\mathbf{x}), \gamma(\mathbf{d}), \mathbf{z}_a, \mathbf{z}_s), \quad (1)$$

where \mathbf{z}_s is the shape code, \mathbf{z}_a is the appearance code, and γ is the positional encoding function to capture high frequency position and view direction details. In our architecture design, the point density $\hat{\sigma}$ is predicted from the positional encoding $\gamma(\mathbf{x})$ conditioned on the shape code \mathbf{z}_s , while the RGB color $\hat{\mathbf{c}}$ additionally depends on the view direction encoding $\gamma(\mathbf{d})$ and the appearance code \mathbf{z}_a .

Volumetric Rendering. Given camera extrinsics ξ , we sample pixel rays $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ given their origin \mathbf{o} and direction \mathbf{d} . We then use the volume rendering operator to calculate the expected color $\hat{\mathbf{c}}(\mathbf{r})$ of the ray \mathbf{r} from the predicted densities and colors (σ_i, c_i) of the sampled points along the ray [27]

$$\begin{aligned} \hat{\mathbf{c}}(\mathbf{r}) &= \sum_{i=1}^N T_i(1 - \exp(-\sigma_i \delta_i)) c_i, \\ \text{where } T_i &= \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \end{aligned} \quad (2)$$

where t_n and t_f represent the near and far bounds, N is the number of evenly-spaced bins between the bounds and $\delta_i = t_{i+1} - t_i$ is the distance of adjacent samples. We predict a soft object foreground mask per ray as $\alpha = 1 - \exp(-\sigma_i \delta_i)$. We follow NeRF [27] to use hierarchical sampling, and calculate the final rendered colors $\hat{\mathbf{c}}(\mathbf{r})$ using coarse and fine points sampled from rays.

3.2. Training

As we lack 3D supervision or multiple views of the in-the-wild images, inspired by some recent approaches [48, 3, 10], we leverage other sources of meta-supervision present in our knowledge of the 3D world.

Input view reconstruction loss. Given the input image, the encoder predicts the 3D object shape and appearance representation, and estimates the pose of the camera that captures the input image. The decoder conditions a NeRF

MLP on the shape and appearance code and uses volume rendering to render an image given a novel view camera pose. We use the decoder to render back the input view by providing it with the predicted camera pose and object representation. In addition to rendering the color image, we render soft object segmentation masks represented by the alpha value $\hat{\alpha}$ in volume rendering. Therefore, for both image reconstruction loss $\mathcal{L}_{\text{recon}}^c$ and mask (alpha) reconstruction loss $\mathcal{L}_{\text{recon}}^\alpha$, we calculate the pixel-wise L_2 loss

$$\mathcal{L}_{\text{recon}}^c = \|\mathbf{c} - \hat{\mathbf{c}}\|_2 \quad (3)$$

$$\mathcal{L}_{\text{recon}}^\alpha = \|\alpha - \hat{\alpha}\|_2 \quad (4)$$

View synthesis adversarial loss. Our method is not provided with novel view images of the objects. To avoid predicting a degenerate 3D representation that only looks like an object from the input view, we utilize adversarial training with a discriminator to force realistic novel view synthesis. During training, a novel view pose ξ' is randomly sampled from a pre-defined prior distribution $p(\xi)$. Then the new view image and alpha mask are rendered and fed together with random input images to a discriminator that ideally shouldn't be able to tell which image is real and which one is fake.

One can think of our model as a generator G^c that receives an image \mathbf{c} , input object mask α and a novel view camera pose ξ' as inputs and generates an image $\hat{\mathbf{c}}'$ as output (internally, there is an encoder inside this generator that encodes the input image into shape and appearance codes, and estimates its camera pose, and a decoder that renders the novel view image conditioned on the predicted codes). We direct the generated images to a traditional convolutional discriminator D^c for adversarial training. We use the following adversarial loss:

$$\mathcal{L}_{\text{adv}}^c = \log D^c(\mathbf{c}) + \log(1 - D^c(G^c(\mathbf{c}, \alpha, \xi'))) \quad (5)$$

Similarly, we use a generator G^α that receives the same inputs but outputs the soft foreground object mask, for which we use a discriminator D^α that tries to distinguish predicted mask from the input image masks. We use the following adversarial loss for the predicted α foreground mask:

$$\mathcal{L}_{\text{adv}}^\alpha = \log D^\alpha(\alpha) + \log(1 - D^\alpha(G^\alpha(\mathbf{c}, \alpha, \xi'))) \quad (6)$$

Pose consistency loss. Our model consists of an encoder and a decoder. The encoder takes in the input image and predicts shape and appearance codes and estimates the pose of the input view camera. Conditioned on the object shape and appearance codes and a randomly sampled camera pose ξ' , the decoder renders a novel view $\hat{\mathbf{c}}'$. We propose a cycle consistency loss that indirectly constrains the estimated

Supervision		chair		car		airplane	
Labeled	Unlabeled	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
100%	0%	18.9	0.77	20.1	0.83	25.2	0.93
1%	99%	18.6	0.77	19.5	0.82	24.8	0.92
1%	0%	17.1	0.72	19.5	0.81	23.7	0.90
0%	100%	16.4	0.72	18.6	0.79	22.2	0.87

Table 1. We perform an ablation study by training our model with different levels of supervision on chair, car, airplane in ShapeNet 3D-R2N2 [5]. This study demonstrates that using only 1% of pose ground truth under the weakly supervised case, and utilizing unlabeled images, leads to performance comparable to fully supervision (100%) of pose.

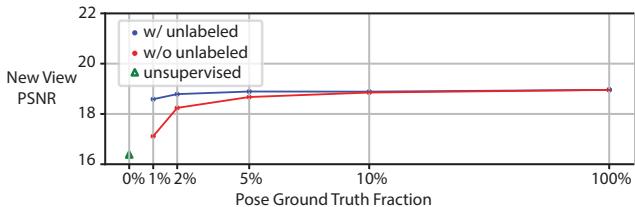


Figure 3. We perform an ablation study by training our model with different levels of pose supervision on ShapeNet 3D-R2N2 [5] chair. Surprisingly, we find using only 1% of pose ground truth under the weakly supervised case, and utilizing unlabeled images, leads to the performance comparable to full supervision (100%).

camera poses to be geometrically meaningful. We feed the rendered novel view \hat{c}' back to the encoder to re-estimate the camera pose $\hat{\xi}'$. We formulate pose consistency loss as

$$\mathcal{L}_{\text{poc}} = \|\xi' - \hat{\xi}'\|_2 \quad (7)$$

Encouraging symmetry. For the categories with symmetric objects, we further force the object shape to be symmetric across the y axis for the objects that have this property. We do this by feeding $x = (x_1, |x_2|, x_3)$ as the input to the neural radiance field MLP. That this only forces symmetry on the object shape. Object appearance additionally takes in the ray direction which lets it to be more flexible.

Pose supervised loss. In our ablation studies, we also study the effect of weakly supervision and full supervision from given pose ground truth, using an additional camera pose estimation loss. The camera pose estimation loss consists of two parts: a mean squared error loss for translation and a mean squared error loss that is applied to the predicted angles in the rotation matrix. Note that such loss is not applied in the unsupervised setting.

4. Experiments

We show quantitative and qualitative results on the ShapeNet 3D-R2N2 [5] dataset. ShapeNet is a synthetic dataset that lets us validate the effectiveness of our method quantitatively and compare it to previous baselines. We present ablation studies that demonstrate the challenges in

learning 3D object representations given only single-view input images with unknown camera poses. Furthermore, in this ablation study, we show the key role that the auxiliary losses and design choices play in helping our model achieve competitive results to a fully-supervised pose setting. Finally, we show qualitative results on the Open Images dataset [19] that demonstrate our method’s capability to learn the 3D presentation of objects and synthesize them from novel views in a large-scale realistic setup in the wild.

4.1. Datasets

We describe the two public benchmark datasets that the models are evaluated on in the following,

ShapeNet 3D-R2N2 [5]. We evaluate our method on three categories including car, chair and airplane. ShapeNet contains 24 views of each object. Since the dataset is synthetic, we could design different sets of ablation studies with providing our model with different levels of supervision, or using the ground truth views to measure the effectiveness of different design choices and losses.

Open Images [19]. Following Ye *et al.* [48], we qualitatively evaluate our method on Open Images dataset. The number of the images in each object category range from 500 to 20k. As presented in [1], the foreground masks are from annotation and are filtered by a fine-tuned occlusion classifier. In our experiments, we crop a tight bounding box around the masked object and use that as the input to our model. In Figure 1 and Figure 7, we show qualitative results on Open Images. We evaluate our model on a wide range of object categories and show that it can learn to represent different types of objects including thin (water tap, guitar, saxophone), flat (hat), round (orange, balloon), deformable (birds, animals), and objects with complicated textures (Christmas trees).

4.2. Experiment Setup

Here we describe the model architecture, training setup and evaluation setup.

Model Architecture. We use an input imagesize of 112×112 . The latent code dimension for both the appearance code z_a and shape code z_s is 64. For both ShapeNet and Open Images datasets, we use pre-defined camera intrinsics (the one used in ShapeNet dataset). For the conditional MLP, we use a depth of 6 with a width of 128. For each ray we use 64 coarse samples and 128 fine samples. We use a near depth of 0.1 and a far depth of 4.0 for the rays. Meanwhile, for ShapeNet dataset, we use a 3D bounding box with the size of $(-0.4, 0.4)$ in x, y, z dimension. The backbone for the encoder is ResNet50 model pre-trained on ImageNet [18]. The backbone for the discriminator consists of 3 fully-convolutional layers with leakyReLU activations and a dropout layer before the last fully connected layer. Since rendering 112×112 rays exceeds the maximum

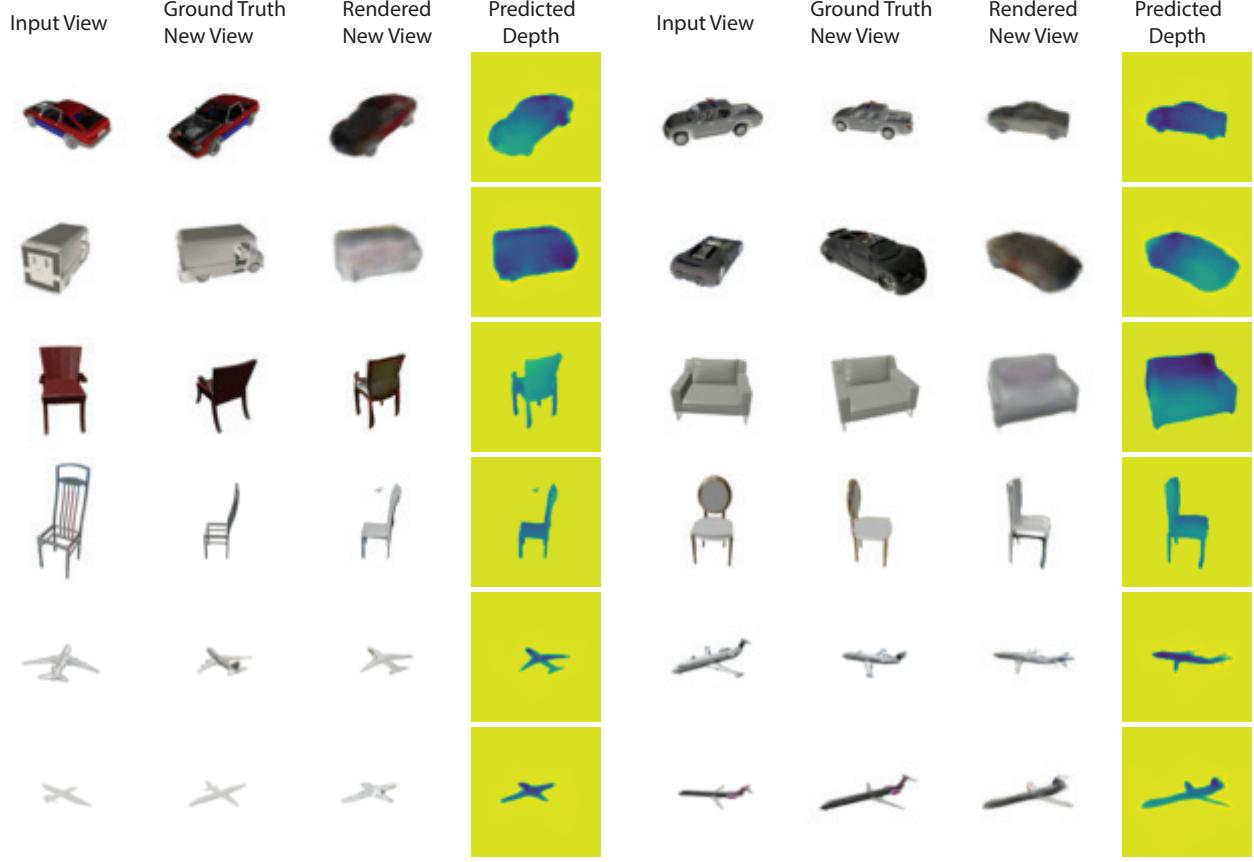


Figure 4. We show our method’s results under the unsupervised setting for chairs, cars and airplanes of the ShapeNet 3D-R2N2 dataset [5].

memory capacity of a GPU, we use a patch discriminator. we stride the image to sample patches of rays during rendering. We use render a sampled novel view patches of size 80×80 for each input image.

Training Setup. During training, the model takes in a single-view masked image as input, and renders back patches for the estimated input view and a randomly sampled novel view. The model is trained with various losses as described in Section 3.2. We use weight 1.0 for input view reconstruction loss $\mathcal{L}_{\text{recon}}^c$ and $\mathcal{L}_{\text{recon}}^\alpha$. We also use weight 1.0 for novel view synthesis adversarial loss $\mathcal{L}_{\text{adv}}^c$ and $\mathcal{L}_{\text{adv}}^\alpha$. We use weight 50.0 for pose consistency loss \mathcal{L}_{poc} .

We use 10 frequencies for positional encoding γ to encourage higher details in rendered images. We find that optimizing pose estimation with positional encoding is difficult. Therefore, similar to BARF [22], we gradually increase the weight of the positional encoding from 0 to 1. Our model is trained for 200 epochs with the Adam optimizer with an initial learning rate of 1^{-3} and a exponential learning rate decay with a rate of 0.96. We train our model on 8 Tesla Nvidia V100 GPUs with a batch size of 8.

Evaluation Setup. Since our model is not provided with ground-truth canonical pose of the objects, the predicted

constraint	recon.	+adv.	+scene box	+pose. consistency
PSNR	12.7	12.8	16.3	16.4
SSIM	0.59	0.64	0.71	0.72

Table 2. We perform an ablation study of losses and design choices such as adversarial loss, scene bounding box, cycle camera pose consistency loss for the unsupervised setting on ShapeNet 3D-R2N2 chair [5]. One additional loss to the objective function or constraint is added at a time (from left to right).

camera poses could be rotated with respect to the ground-truth. This is still a valid prediction since the object representation is also rotated with respect to the ground-truth canonical pose. To evaluate the accuracy of our rendered novel view images quantitatively, we use a universal offset matrix during evaluation that maps the predicted poses to the ground-truth ones.

4.3. Ablation Study

As shown in Figure 3 and Table 1, we inspect our model under different levels of pose ground truth supervision (with single-view input). We show that our model by using only 1% of posed data and leveraging the other 99% of unlabeled



Figure 5. We perform an ablation study of losses and design choices such as adversarial loss, scene bounding box, cycle camera pose consistency loss for the unsupervised setting on ShapeNet 3D-R2N2 chair [5]. One additional loss to the objective function or constraint is added at a time (from left to right).

data, could achieve results comparable to a fully supervised setting, and outperform training without unlabeled data.

Meanwhile, we also perform studies with various combinations of losses and design choices for the unsupervised condition: adversarial loss \mathcal{L}_{adv} , scene box, and pose consistency loss \mathcal{L}_{poc} . We find that using photometric reconstruction loss ($\mathcal{L}_{\text{recon}}^c$) only does not converge in our problem setup where we only have access to single-view images with unknown camera poses. We demonstrate the key role each of the following losses or design choices make in enabling our method achieve competitive results: (1) adversarial loss for images $\mathcal{L}_{\text{adv}}^c$ and masks $\mathcal{L}_{\text{adv}}^\alpha$, (2) scene box, and (3) cycle camera pose consistency loss \mathcal{L}_{poc} . Results are shown in Figure 5 and Table 2. Furthermore, we show qualitative results in other categories including car and airplane from ShapeNet dataset in Figure 4.

4.4. Evaluation on Benchmark Datasets

Results on ShapeNet 3D-R2N2. Meanwhile, we compare our model im2nerf with other baselines in the unsupervised settings using single-view and unknown camera pose, our proposed model im2nerf achieves significantly better results in comparison to the baselines on the ShapeNet 3D-R2N2 dataset [5], as shown in Table 3.

- *Shelf-Supervised.* We compare our method to that of Ye *et al.* [48]. They use a combination of voxel and implicit representations to predict the 3D shape and appearance of the objects. We refer to this model as *shelf-supervised* in the our results. In order to make the comparison, We use their publicly released source code to perform the evaluations on both ShapeNet 3D-R2N2 and Open Images datasets. For Open Images dataset, we use their released pre-trained model for each category to render the images.
- *PrGAN.* We compare our method to that of Gadelha *et al.* [7] which we call PrGAN throughout the paper. We use the open source implementation by Ye *et al.* [48] to evaluate PrGAN on our datasets.
- *HoloGAN.* We compare our method to that of Nguyen *et al.* [29]. HoloGAN is added with a function that

model	HoloGAN	PrGAN	Shelf-Supervised	im2nerf (ours)
chair	13.2	13.8	13.9	16.4
car	16.2	16.7	16.9	18.6
airplane	16.0	16.3	16.4	22.2

Table 3. We compare our method quantitatively with baselines including Shelf-Supervised [48], HoloGAN [29], PrGAN [7] evaluated on the unsupervised condition for ShapeNet [5] chair, car, and airplane. We use the average PSNR of the synthesized novel view images as the comparison metric.



Figure 6. Our method im2nerf’s qualitative results on Open Images dataset [19]. The synthesized novel views by our algorithm look more realistic than those of the Ye *et al.* [48].

maps the volumetric feature to occupancy using reprojection consistency with foreground masks. We use the open source implementation by Ye *et al.* [48] to evaluate HoloGAN on our datasets.

Results on Open Images dataset. We show qualitative results on Open Images dataset. In Figure 6, we compare our results qualitatively with that of Ye *et al.* [48]. In Figure 1 and Figure 7, we show qualitative results of our proposed method im2nerf on Open Images dataset.

5. Conclusion and Limitations

In this paper, we present im2nerf, a learning framework that can decompose an image of an object into a 3D shape code, an appearance code, and a camera viewpoint. Our



Figure 7. Results of im2nerf model evaluated on Open Images dataset [19, 1]. We show three columns of results, where in each column, from left to right we show the input image, synthesized novel view, and predicted depth of the synthesized novel view.

model conditions a NeRF on the 3D shape and appearance codes and uses volume rendering to synthesize novel views of the object. We exploit the meta-supervision that is present in our knowledge of the 3D world to solve this highly under-constrained problem. In particular, in addition to using a reconstruction loss on the synthesized input view, we use an auxiliary adversarial loss on rendered novel view, leverage the fact that most objects are symmetric, and apply a scene box and cycle pose consistency loss. Interestingly, we find using only 1% of pose ground truth under the weakly supervised case, and utilizing unlabeled images, leads to performance comparable to full supervision. We will release code and checkpoints upon acceptance.

While we achieve encouraging results across both syn-

thetic and large-scale real-world datasets, our approach has several limitations. Conditional NeRF consumes a significant amount of memory and computation, which limits our ability to render high-resolution novel view images during training. Therefore, we apply the adversarial loss on coarse patches rendered from the novel view. We believe this causes our model to learn a less detailed 3D representation. Additionally, since we are using single-view images, the model may have trouble learning concave structures. To solve this issue, we believe either the model should be provided with a pseudo-depth supervision or alternatively another auxiliary self-supervised loss that would encourage the model to learn shape from shading.

References

- [1] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *CVPR*, 2019. 5, 8, 11
- [2] Mark Boss, Raphael Braun, Varun Jampani, Jonathan Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural Reflectance Decomposition from Image Collections. In *ICCV*, October 2021. 2
- [3] Eric Chan, Marco Monteiro, Peter Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis. In *CVPR*, 2021. 1, 2, 3, 4
- [4] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *International Conference on 3D Vision (3DV)*, 2016. 2
- [5] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5, 6, 7, 11, 12
- [6] Terrance DeVries. Unconstrained Scene Generation With Locally Conditioned Radiance Fields. In *ICCV*, October 2021. 2
- [7] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *2017 International Conference on 3D Vision (3DV)*, pages 402–411. IEEE, 2017. 3, 7, 11
- [8] Rohit Girdhar, David F. Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2
- [9] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, 2019. 2
- [10] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 1, 3, 4
- [11] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2
- [12] Philipp Henzler. Unsupervised Learning of 3D Object Categories from Videos in the Wild. In *CVPR*, 2021. 2
- [13] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *ICCV*, October 2021. 1, 2
- [14] Wongbong Jang and Lourdes Agapito. CodeNeRF: Disentangled Neural Radiance Fields for Object Categories. In *ICCV*, October 2021. 1, 2, 4
- [15] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *CVPR*, 2020. 2, 3
- [16] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 2, 3
- [17] Adam Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Rezende. NeRF-VAE: A Geometry Aware 3D Scene Generative Model. In *ICML*, 2021. 2, 4
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 5
- [19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallochi, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 5, 7, 8, 11
- [20] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. In *arXiv:2103.02597*, 2021. 2
- [21] Tzu-Mao Li, Miika Aittala, Fredo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. In *ACM TOG*, 2018. 2, 3
- [22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-Adjusting Neural Radiance Fields. In *ICCV*, October 2021. 1, 2, 6
- [23] Shichen Liu, Tianye Li, Weikai Chen, , and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019. 2, 3
- [24] Ricardo Martin-Brualla, Noha Radwan, Mehdi Sajjadi, Jonathan Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 2
- [25] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. *arXiv preprint arXiv:2103.15606*, 2021. 2
- [26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*, 2019. 2
- [27] Ben Mildenhall, Pratul Srinivasan, Matthew Tancik, Jonathan Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*, pages 405–421. Springer, 2020. 1, 2, 4
- [28] Mahyar Najibi, Guangda Lai, Abhijit Kundu, Zhichao Lu, Vivek Rathod, Thomas Funkhouser, Caroline Pantofaru, David Ross, Larry S. Davis, and Alireza Fathi. Dops: Learning to detect 3d objects and predict their 3d shapes. In *CVPR*, 2020. 2
- [29] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7588–7597, 2019. 2, 3, 7, 11
- [30] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *CVPR*, 2021. 1, 2, 3
- [31] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable Volumetric Rendering: Learning Implicit 3D Representations without 3D Supervision. In *CVPR*, 2020. 2, 3

- [32] Jeong Joon Park, Pete Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *CVPR*, 2019. 2, 3
- [33] 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. 1
- [34] Konstantinos Rematas and Vittorio Ferrari. Neural Voxel Renderer: Learning an Accurate and Controllable Rendering Tool. In *CVPR*, 2020. 2
- [35] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis. In *Adv. Neural Inform. Process. Syst.*, 2020. 3
- [36] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 2
- [37] Karl Stelzner, Kristian Kersting, and Adam Kosiorek. Decomposing 3D Scenes into Objects via Unsupervised Volume Segmentation. <https://arxiv.org/abs/>, 2021. 2
- [38] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *CVPR*, 2018. 2
- [39] Alex Trevithick and Bo Yang. GRF: Learning a General Radiance Field for 3D Scene Representation and Rendering. In *ICCV*, October 2021. 2
- [40] Shubham Tulsiani, Saurabh Gupta, David Fouhey, Alexei A. Efros, and Jitendra Malik. Factoring shape, pose, and layout from the 2d image of a 3d scene. In *CVPR*, 2018. 2
- [41] GÜl Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2
- [42] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*, 2021. 1, 2
- [43] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF-: Neural Radiance Fields Without Known Camera Parameters. <https://arxiv.org/abs/2102.07064>, 2021. 2
- [44] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative adversarial modeling. In *NeurIPS*, 2016. 2
- [45] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 1, 3
- [46] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Lin-guang Zhang, Xiaou Tang, , and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 2
- [47] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014. 2
- [48] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8843–8852, 2021. 1, 3, 4, 5, 7, 11
- [49] Lin Yen-Chen, Pete Florence, Jonathan Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In *IROS*, 2021. 2
- [50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*, 2021. 2, 11
- [51] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-Place Scene Labelling and Understanding with Implicit Scene Representation. In *ICCV*, October 2021. 1
- [52] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, 2019. 2

A. Evaluation Details

In this section we clarify the implementation details presented in the paper.

Weakly-Supervision. The results demonstrated in Table 1, Figure 3 in the main paper and Figure 8, are evaluated on ShapeNet 3D-R2N2 dataset [5]. In those experiments, single-view images are used as input. In the weakly supervised setup, a small fraction of those single-view images are provided with ground-truth camera pose. And we first pre-train the model with images that have known camera pose. Then we fine-tune using both posed images as well as images with unknown camera pose.

As shown in Figure 8, we inspect our model under different levels of pose ground truth supervision (with single-view input). We show that our model by using only 1% of posed data and leveraging the other 99% of unlabeled data, could achieve results comparable to a fully supervised setting, as well as outperform training without unlabeled data.

Unsupervised setting. In the unsupervised setup, none of the images is provided with ground-truth camera pose. In these experiments, camera pose is predicted from the input single-view image. And during training, our model does not have access to ground-truth camera pose. The results of our model with no pose supervision on ShapeNet 3D-R2N2 dataset [5] are demonstrated in Table 2, Table 3, Fig 4, Fig 5. The unsupervised results on Open Images dataset [19, 1] are shown in Fig 6, Fig 7.

Multi-view Supervision. We also evaluate our model in a setting where the pose is given as input, and the model has access to multiple pose views of an object during training (during evaluation it is still provided with a single-view images).

In the Table A below, We summarize our results for these different settings.

B. Comparison to Baselines

As described in the paper, we compare our method with multiple baselines: PrGAN [7], HoloGAN [29] and shelf-supervised mesh prediction by Ye et al [48]. We use the open-source implementation of PrGAN and HoloGAN by Ye et al [48]. HoloGAN is added with a function that maps the volumetric feature to occupancy using re-projection consistency with foreground masks. Each model only takes a single-view input and renders new views. These models are also trained with single-view unposed images using meta-supervision. Our model outperforms these approaches to a large extent, as demonstrated in Table 3 and Figure 6 in the main paper.

Meanwhile, we also implement pixelnerf [50], which uses a spatial encoder to generate image feature as local embedding to the decoder, while model uses global latent code instead. For the local embedding, pose are required

as inputs to project rays onto the image features during rendering, we find training such model with a spatial encoder without pose given is very difficult and the results of using local embedding are much worse than those from our models using global latent code.

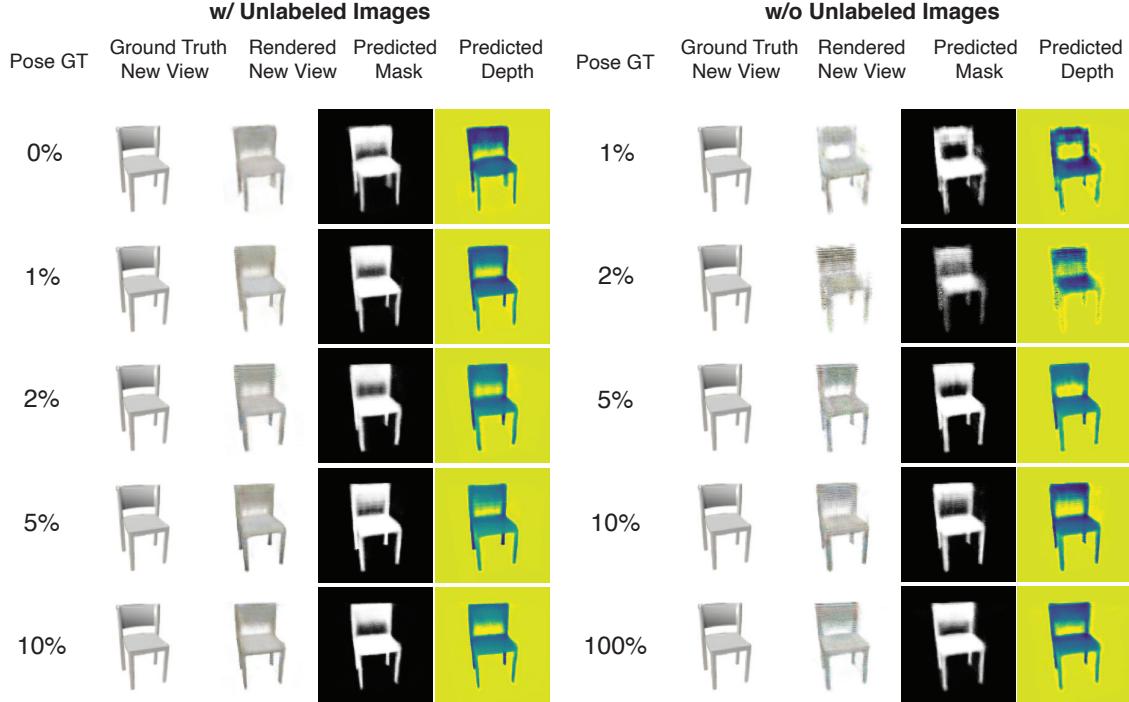


Figure 8. Visualization of our results with different levels of supervision on ShapeNet 3D-R2N2 [5] chair. We find out that using as little as 1% posed data, we can leverage the remaining 99% unlabeled data to achieve results that are comparable to fully supervision (100%) of pose.

View	Category	Chair		Car		Airplane	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
multi-	input	20.8	0.83	22.2	0.86	25.8	0.94
	input	19.3	0.80	21.0	0.85	25.4	0.93
single-	fully-supervised	19.0	0.77	20.1	0.83	25.2	0.93
	weakly-supervised	18.6	0.77	19.5	0.82	24.8	0.92
	unsupervised	16.4	0.72	17.2	0.74	22.2	0.87

Table 4. Comparison between model with multi-view v.s. single-view supervision, and whether pose is directly given as input or predict from the single-view image input given different levels of supervision (fully-supervised 100% v.s. weakly-supervised 1% v.s. unsupervised 0%) on ShapeNet 3D-R2N2 dataset [5] chair, car and airplane.