😂LOLNeRF: Learn from One Look

Daniel Rebain [1,3] Mark Matthews[3] Kwang Moo Yi[1] Dmitry Lagun[3]
Andrea Tagliasacchi[2,3]
[1]University of British Columbia [2]Simon Fraser University [3]Google Research

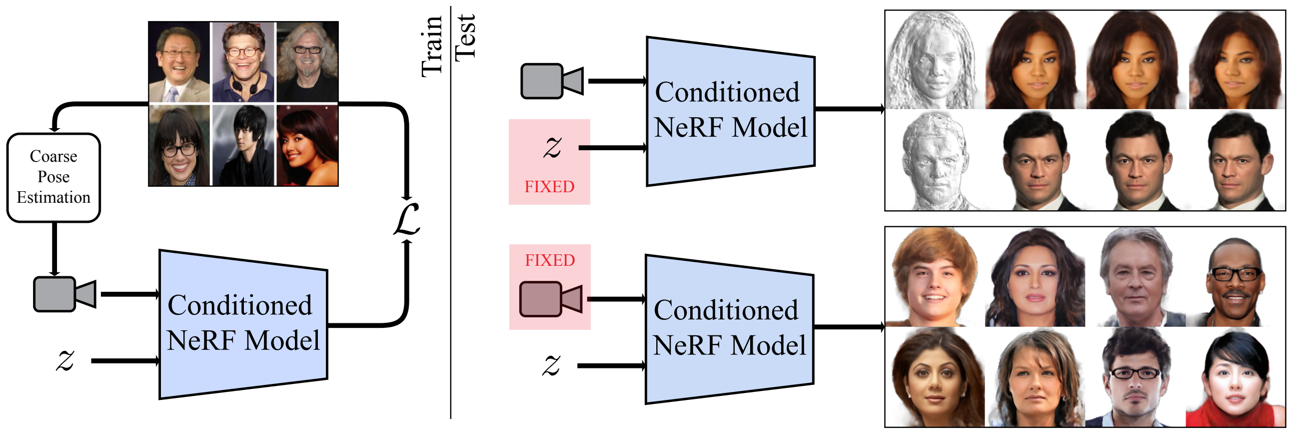

Figure 1. **Teaser** – Our method learns a space of shape and appearance by reconstructing a large collection of *single-view* images using a single neural network conditioned on a *shared* latent space (left). This allows for a volumetric 3D model to be lifted from the image, and rendered from novel viewpoints (right). https://lolnerf.github.io

Abstract

We present a method for learning a generative 3D model based on neural radiance fields, trained solely from data with only single views of each object. While generating realistic images is no longer a difficult task, producing the corresponding 3D structure such that they can be rendered from different views is non-trivial. We show that, unlike existing methods, one does not need multi-view data to achieve this goal. Specifically, we show that by reconstructing many images aligned to an approximate canonical pose with a single network conditioned on a shared latent space, you can learn a space of radiance fields that models shape and appearance for a class of objects. We demonstrate this by training models to reconstruct object categories using datasets that contain only one view of each subject without depth or geometry information. Our experiments show that we achieve state-of-the-art results in novel view synthesis and high-quality results for monocular depth prediction. https://lolnerf.github.io

1. Introduction

A long-standing challenge in computer vision is the extraction of 3D geometric information from images of the real world [37]. Understanding 3D geometry is critical to understanding the physical and semantic structure of objects and scenes, but achieving it remains a very challenging problem. Work in this area has mainly focused either on deriving geometric understanding from more than one view [1, 25, 62], or by using known geometry to supervise the learning of geometry from single views [10, 14, 18, 43]. Here, we take a more ambitious approach and aim to derive equivalent 3D understanding in a generative model from only *single* views of objects, and without relying on explicit geometric information like depth or point clouds. Deriving such 3D understanding is, however, non trivial. While Neural Radiance Field (NeRF)-based methods [13, 44] have shown great promise in geometry-based rendering, they focus on learning a single scene from multiple views.

Existing NeRF works [16, 44, 52] all require supervision from more than one viewpoint, as without it, NeRF methods are prone to collapse to a flat representation of the scene, because they have no incentive to create a volumetric representation; see Figure 2 (left). This serves as a major bottleneck, as multiple-view data is hard to acquire. Thus, architectures have been devised to workaround this that combine NeRF and Generative Adversarial Networks (GANs) [9, 47, 57], where the multi-view consistency is enforced through a discriminator to avoid the need for multi-view training data.

In this work we show that – surprisingly – having only single views of a class of objects is enough to train NeRF models without adversarial supervision, as long as a *shared* generative model is trained, and *approximate* camera poses are provided. In a nutshell, the multi-view constraint of existing works no longer necessarily needs to be enforced, and cameras do not have to be accurate to achieve compelling results; see Figure 2 (right). Specifically, we roughly align all images in the dataset to a canonical pose using predicted 2D landmarks, which is then used to determine from which view the radiance field should be rendered to reproduce the original image. For the generative model we employ an auto-decoder framework [51]. To improve generalization, we further train two models, one for the foreground – the common object class of the dataset – and one for the background, since the background is often inconsistent throughout the data, hence unlikely to be subject to the 3D-consistency bias. We also encourage our model to model shapes as solid surfaces (i.e. sharp outside-to-inside transitions), which further improves the quality of predicted shapes; see the improvements from Figure 2 (middle) to Figure 2 (right).

A noteworthy aspect of our method is that we *do not* require rendering of entire images, or even patches, while training. In the auto-decoder framework, we train our models to reconstruct images from datasets, and at the same time find the optimal latent representations for each image – an objective that can be enforced on *individual* pixels. Hence, our method can be trained with arbitrary image sizes without any increase in memory requirement during training. In contrast, existing methods that utilize GANs [9,47,57] supervise on inter-pixel relationships through their discriminators, greatly limiting or outright preventing them from being able to scale with respect to training image resolution. In summary, we:

Figure 2

- propose a method for learning 3D reconstruction of object categories from *single-view* images which decouples training complexity from image resolution;
- show that single views are enough to learn high-quality prediction of geometry (e.g. depth) without any geometric supervision (Figure 3);
- show that our method exceeds adversarial methods in representing appearance of objects from the learned category by reconstructing held-out images and novel views.

2. Related Work

There are three main research topics related to ours: (1) classic 3DMMs, (2) neural implicit shape representations, and (3) shape estimation from single view. We also review NeRF and auto-decoders, on which we base our research.

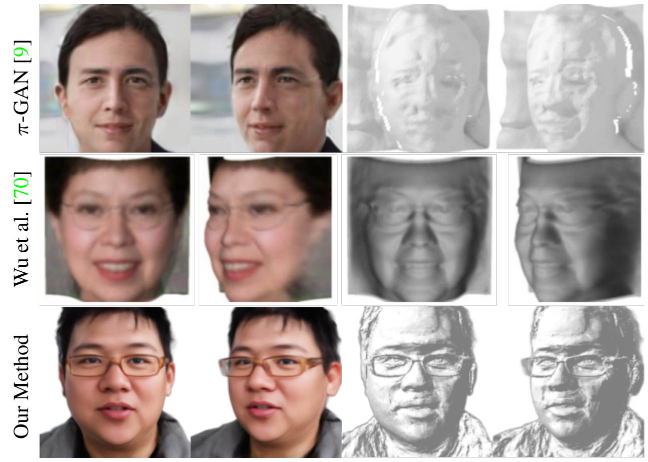

Figure 3. **Geometry recovery** – Our model is able to capture reasonable coarse geometry from a single view as well as topologically complex detail like glasses that can't be encoded in a depth map.

3D Morphable Models (3DMM). Classical approaches to building shape spaces have focused on parameterized 3D mesh representations, with much work devoted to human faces [5,19,54], as surveyed by Egger et al. [15]. These models are typically built from geometric information sources like 3D scans or depth sensors, which are highly accurate, but require significant investment (e.g. the proprietary Disney Medusa capture rig [20,55] or the commercially available solutions provided by 3DMD [66]). The use of mesh representations also limits these models for applications like novel view synthesis where phenomena like hair are not well reproduced. By contrast, our approach is relatively unconstrained, as it derives flexible volumetric representations directly from images taken in uncontrolled environments.

Neural Implicit Representations. Representing scenes as 3D implicit fields has proven successful for a number of tasks. These models can take a number of forms, including representations of distance [51], occupancy [11,43], learned scene features [62], and light fields [61]. One representation in particular has been highly successful in simultaneously modeling shape and appearance: radiance fields. Neural Radiance Field models, known as NeRFs [44], use fields for density and radiance, and are particularly effective in learning 3D scene structure from images alone. A substantial number of extensions to NeRF have been proposed [13], some notable examples being: per-view appearance codes [39], multi-resolution training [3], camera co-optimization [35, 47, 69], hard surface priors [49], deformable scenes [52], variable topology [53] and foreground-background decompositions [74]. The single-scene formulation of NeRF has also been extended to general object classes with [25, 71], and hybridized with GAN methods as in GNeRF [42], GIRAFFE [48], and StyleNeRF [22].

Shape from Single View. A long-standing objective in computer vision has been to understand the 3D structure of scenes and objects from a single image. Many works have approached this problem by encoding a relationship between appearance and structure using prior knowledge of that structure as supervision [14, 18, 23, 43]. Until recently however, the problem of deriving such a model from only single-view observations has remained very difficult. Wu et al. [70] demonstrate how shape can be inferred for categories of objects which are approximately symmetric.

The most relevant to our work however, has been the development of GAN-based methods which learn a space of shapes that when rendered produce a distribution of images is indistinguishable from a training distribution [9, 47, 50, 57]. This approach is effective in producing models with plausible structure, as they enforce an implicit multi-view constraint by requiring that renders from any viewpoint appear realistic. Unfortunately, this requires the use of discriminator networks which are very inefficient when combined with 3D representations that use volumentric representations. To avoid this limitation, we reconstruct images directly with a more efficient and scalable stochastic sampling process.

Neural Radiance Fields (NeRF). Neural Radiance Fields [44] use classical volume rendering [28] to compute radiance values for each pixel $\mathbf{p}$ from samples taken at points $\mathbf{x}$ along the associated ray. These samples are computed using a learned *radiance field* which maps $\mathbf{x}$, as well as the ray direction $\mathbf{d}$, to radiance values c and density values σ. The volume rendering equation takes the form a weighted sum of the radiance values at each sample point $\mathbf{x}_i$:

$$C_{\text{NeRF}}(\mathbf{p}) = \sum_{i=1}^{N} w_i \cdot c(\mathbf{x}_i, \mathbf{d}) \tag{1}$$

with the weights w_i being derived from an accumulation of the transmittance along the view ray $\mathbf{x}_i$:

$$w_i = (1 - \exp(-\sigma(\mathbf{x}_i)\delta_i)) \cdot \exp\left(-\sum_{j=1}^{i-1} \sigma(\mathbf{x}_j)\delta_j\right) \tag{2}$$

where δ_i is the sample spacing at the i-th point. Note here that we denote the product of the accumulated transmittance and sample opacity as w, as this value determines the contribution of a single sample to the final pixel value. These weights can also be used to compute other values such as surface depth (by replacing the per sample radiance values with sample depth $d(\mathbf{x}_i)$, or the overall pixel opacity:

$$D(\mathbf{p}) = \sum_{i=1}^{N} w_i \cdot d(\mathbf{x}_i) \qquad \alpha(\mathbf{p}) = \sum_{i=1}^{N} w_i \cdot 1 \tag{3}$$

Auto-decoders. Auto-decoders [51, 60, 64], also known as Generative Latent Optimization (GLO) [2, 6, 39, 51, 59], are a family of generative models that learn without the use of either an encoder or discriminator. The method works similarly to an auto-*encoder*, in that a decoder network maps a latent code to a final output. However the method differs in how these latent codes are found: auto-decoders learn the codes directly by allocating a table of codes with a row for each distinct element in the training dataset. These codes are *co-optimized* with the rest of the model parameters as learnable variables.

Figure 4. **Novel view synthesis** – For each method we show an example appropriate to the training method: for π-GAN [9] a latent code sampled from the training distribution, and for ours a learned latent code which reconstructs a training image. Our method recovers more sharp detail due to training on higher resolution images. For a comparison of novel views of the same image reconstructed by both methods, see Figure 8.

3. Method

We visualize our architecture in Figure 5. We train our network parameters and latent codes $\mathbf{Z}$ by minimizing the weighted sum of three losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rgb}} + \underbrace{\lambda_{\text{mask}}\mathcal{L}_{\text{mask}}}_{\text{Section 3.1}} + \underbrace{\lambda_{\text{hard}}\mathcal{L}_{\text{hard}}}_{\text{Section 3.2}} \tag{4}$$

where the first term is the standard L2 photometric reconstruction loss over pixels $\mathbf{p}$ from the training images I_k:

$$\mathcal{L}_{\text{rgb}} = \mathbb{E}_{k\in\{1..K\},\mathbf{p}\in I_k}\left[(C(\mathbf{p}|\mathbf{z}_k) - C^k_{\text{GT}}(\mathbf{p}))^2\right] \tag{5}$$

We extend the "single-scene" (i.e. overfitting/memorization) formulation of NeRF to support learning a *latent space of shapes* by incorporating an auto-decoder architecture. In this modified architecture, the main NeRF backbone network is conditioned on a per-object latent code $\mathbf{z} \in \mathbb{R}^D$, as well as the L-dimensional positional encoding $\gamma^L(x)$ as in [44]. Mathematically, the density and radiance functions are then of the form $\sigma(\mathbf{x}|\mathbf{z})$ and $c(\mathbf{x}|\mathbf{z})$; note we consider a formulation where radiance is not a function of view direction $\mathbf{d}$. These latent codes are rows from the latent table $\mathbf{Z}\in\mathbb{R}^{K\times D}$, which we initialize to $\mathbf{0}^{K\times D}$, where K is the number of images. This architecture makes it possible to accurately reconstruct training examples without requiring significant extra computation and memory for an encoder model, and avoids requiring a convolutional network to extract 3D information from the training images [67, 72]. Training this model follows the same procedure as single-scene NeRF, but draws random rays from all K images in

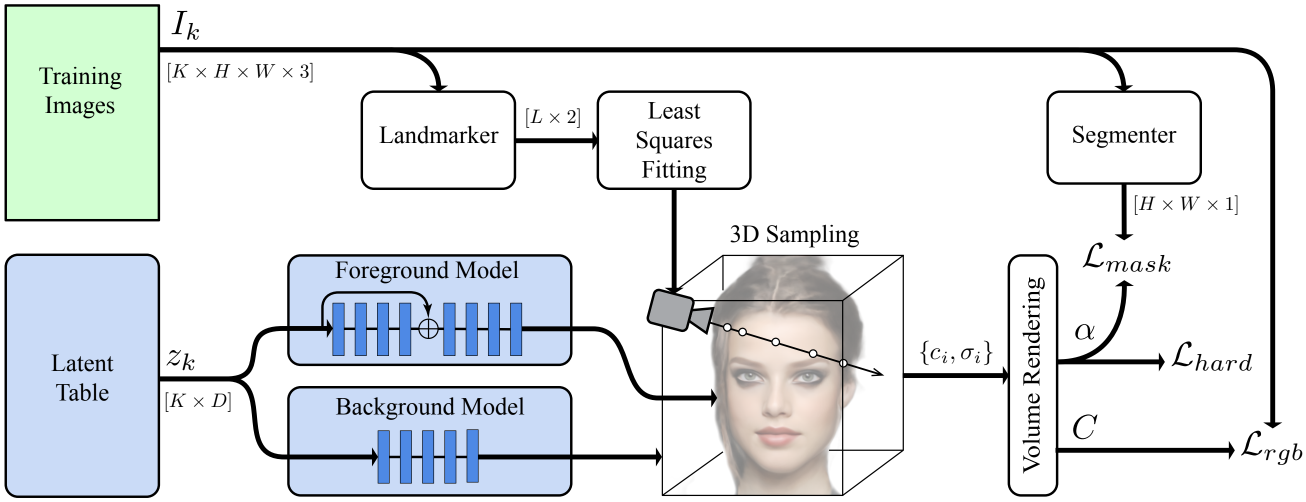

Figure 5. **Architecture** – Our method learns a per-image table of latent codes alongside foreground and background NeRFs. Volumetric rendering output is subject to a per-ray RGB loss against each training pixel, and alpha value against an image segmenter. Camera alignments are derived from a least-squares fit of 2D landmarker outputs to class-specific canonical 3D keypoints.
Legend: input data learned fixed function

the dataset, and associates each ray with the latent code that corresponds to the object in the image it was sampled from.

3.1. Foreground-Background Decomposition

Similar to [17, 47, 74], we use a separate model to handle the generation of background details. We use a lower-capacity model $C_{\text{bg}}(\mathbf{d}|\mathbf{z})$ for the background that predicts radiance on a per-ray basis. We then render by combining the background and foreground colors using a transparency value derived from the NeRF density function:

$$C(\mathbf{p}|\mathbf{z}) = \alpha(\mathbf{p}|\mathbf{z}) \cdot C_{\text{NeRF}}(\mathbf{p}|\mathbf{z}) + (1 - \alpha(\mathbf{p}|\mathbf{z})) \cdot C_{\text{bg}}(\mathbf{d_p}|\mathbf{z}) \quad (6)$$

In practice supervising the foreground/background separation is not always necessary; see the SRN Cars [62] results in Figure 10 which learn a foreground decomposition naturally from solid background color and 360° camera distribution. When a pre-trained module is available for predicting the *foreground segmentation* of the training images, we also apply an additional loss to encourage the transparency of the NeRF volume to be consistent with this prediction:

$$\mathcal{L}_{\text{mask}} = \mathbb{E}_{k \in \{1..K\}, \mathbf{p} \in I_k} \left[(\alpha(\mathbf{p}|\mathbf{z}_k) - S_I(\mathbf{p}))^2 \right] \quad (7)$$

where $S_I(\cdot)$ is the pre-trained image segmenter applied to image I_k and sampled at pixel $\mathbf{p}$. When training on face datasets, we employ the *MediaPipe* Selfie Segmentation [41] for the pre-trained module in (7) and λ_{mask}=1.0.

3.2. Hard Surfaces

NeRF does not explicitly enforce that the learned volumetric function strictly model a hard surface. With enough input images, and sufficiently textured surfaces, multi-view consistency will favor the creation of hard transitions from empty to solid space. Unfortunately, this property does not hold in the single view case. Because the field function that corresponds to each latent code is only supervised from *one* viewpoint, this often results in blurring of the surface *along* the view direction; see Figure 2. To counter this, we impose a prior on the probability of the weights w to be distributed as a mixture of Laplacian distributions, one with mode around weight zero, and one with mode around weight one:

$$\mathbb{P}(w) \propto e^{-|w|} + e^{-|1-w|} \quad (8)$$

Note the distribution is peaky, and will encourage a sparse solution where any values of w in the open interval $(0, 1)$ to be discouraged. We convert this prior into a loss via:

$$\mathcal{L}_{\text{hard}} = -\log(\mathbb{P}(w)) \quad (9)$$

The magnitude of $\sigma(\mathbf{x})$ which satisfies this constraint depends on the sampling density. Equation (9) encourages the density to produce a step function that saturates sampling weight over at least one sampling interval, which, by construction, is appropriate for the scale of scene being modelled. We employ $\lambda_{\text{hard}} = 0.1$ in our experiments.

3.3. Camera Parameters

Volume rendering requires camera parameters that associate each pixel with a ray used to compute sample locations. In classic NeRF, cameras are estimated by structure-from-motion on the input image dataset. For our single-view use case, this is not possible due to depth ambiguity. To make our method compatible with single-view images, we we employ the *MediaPipe* Face Mesh [40] pre-trained network

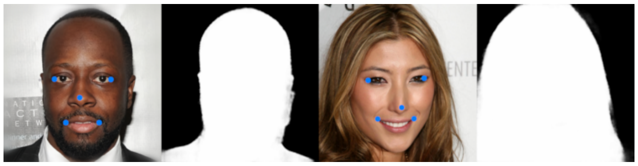

Figure 6. **Landmarker and Segmenter** – (Samples outputs from the landmarker and the segmenter networks for two input identities. Blue circles (●) denote the identified landmarks.

module to extracts 2D landmarks that appear in consistent locations for the object class being considered. Figure 6 shows example network output of the five landmarks used for human faces.

These landmark locations are then aligned with projections of canonical 3D landmark positions with a “shape matching” least-squares optimization to acquire a rough estimate of camera parameters; see the supplementary material for more detail on this process.

3.4. Conditional Generation

Given a pre-trained model, we can find a latent code z which reconstructs an image which was not present in the training set. As the latent table is learned in parallel with the NeRF model parameters, we can treat this process as a fine-tuning optimization for an additional row in the latent table. This row is initialized to the mean $\mu_{\mathbf{Z}}$ over the existing rows of the latent table, and is optimized using the same losses and optimizer as the main model. Results for this fitting method are shown in Section 4.

3.5. Unconditional Generation

To sample novel objects from the space learned by our model, we sample latent codes from the empirical distribution $\mathcal{Z}$ defined by the rows of the latent table $\mathbf{Z}$. We model $\mathcal{Z}$ as a multivariate Gaussian with mean $\mu_{\mathbf{Z}}$ and covariance $\chi_{\mathbf{Z}}$ found by performing principal component analysis on the rows of $\mathbf{Z}$. Similar to other generative models which use a Gaussian prior on their latent variables, we observe a trade-off between diversity and quality of samples when sampling further away from the mean of the distribution. Thus, we employ the “truncation trick” commonly used in the GAN literature [8, 30, 32, 38] to control this trade-off.

4. Results

We visualize images rendered from our models trained on the CelebA-HQ [29], FFHQ [30], AFHQ [12], and SRN Cars [62] datasets in Figure 10. We include additional qualitative results from these datasets in the supplementary material. To provide quantitative evaluation of our method and comparison to state of the art we perform a number of experiments, described in the following subsections.

Method	PSNR↑	SSIM↑	LPIPS↓	Res.
π-GAN [9] (CelebA)	23.5	0.858	0.226	
Ours (FFHQ)	29.0	0.913	0.199	128^2
Ours (CelebA-HQ)	**29.1**	**0.914**	**0.197**	

Table 1. **Reconstructions of training images** – Metrics on a subset of 200 images from the π-GAN [9] training set. Our model achieves significantly higher reconstruction quality, regardless of whether it is trained on (FFHQ) or (CelebA-HQ).

Method	PSNR↑	SSIM↑	LPIPS↓	Res.
π-GAN [9] (CelebA)	21.8	0.796	0.412	256^2
Ours (CelebA-HQ)	**26.2**	**0.856**	**0.363**	
π-GAN [9] (CelebA)	20.9	0.795	0.522	
Ours (CelebA-HQ)	25.1	0.831	0.501	512^2
Ours (FFHQ)	**25.3**	**0.836**	**0.491**	

Table 2. **Reconstructions of test images** – Reconstruction quality (rows 1 and 2) of models trained on (CelebA) and (CelebA-HQ) on images from a 200-image subset of FFHQ, and (rows 3-5) of models trained at 256^2 (Ours) and 128^2 (π-GAN [9]) on high resolution 512^2 versions of the test images.

4.1. Image Reconstruction – Table 1 and Table 2

As LOLNeRF is trained with an image reconstruction metric, we first perform experiments to evaluate how well images from the training dataset are reconstructed. In Table 1, we show the average image reconstruction quality of both our method and π-GAN [9] for a 200-image subset of the π-GAN [9] training set (CelebA [36]), as measured by peak signal to noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). To compare against π-GAN [9], which does not learn latent codes corresponding to training images, we use the procedure included with the original π-GAN [9] implementation for fitting images through test-time latent optimization. Because this assumes perfectly forward facing pose, to make the comparison fair, we augment it with our camera fitting method to improve its results on profile-view images. We also perform a more direct comparison of image fitting by testing on a set of held out images *not seen* by the network during training. To do this, we sample a set of 200 images from the FFHQ dataset and use the latent optimization procedure described in Section 3.4 to produce reconstructions using a model trained on CelebA images. We show the reconstruction metrics for these images using LOLNeRF and π-GAN [9] in Table 2 and examples of the reconstructed images in Figure 7.

4.2. Novel View Synthesis – Table 3

To evaluate the accuracy of the learned 3D structure, we perform image reconstruction experiments for synthesized

Figure 7. **Conditional generation** – we show examples of both our method (trained on CelebA-HQ@256^2) and π-GAN (trained on CelebA@128^2) fitting to images not seen during training. Our method produces much sharper reconstructions, especially for non-frontal views and appearances that are not well represented in the training set.

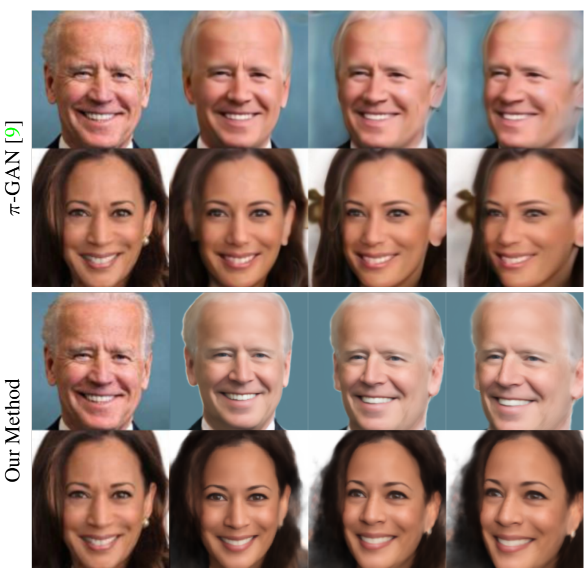

Figure 8. **Novel view comparison** – Comparison of novel views on held-out image fitting with our method and π-GAN [9]. Our method captures more fine detail and produces sharper results at large angles.

novel views. We render these novel views by performing image fitting on single frames from a synchronized multi-view face dataset, Human Multiview Behavioural Imaging (HUMBI) [73], and reconstructing images using the camera parameters from other ground truth views of the same person. The results of this experiment for LOLNeRF and π-GAN [9] are given in Table 3. We find that our model achieves significantly better reconstruction from novel views, indicating that our method has indeed learned a better 3D shape space than π-GAN [9] – a shape space that is capable of generalizing to unseen data and does more than simply reproducing the query image from the query view. We also show qualitative examples of novel views rendered by LOLNeRF and π-GAN [9] in Figures 4 and 8.

Method	PSNR↑	SSIM↑	Mask-LPIPS↓
π-GAN [9] (CelebA)	24.5	0.918	0.102
Ours (CelebA-HQ)	26.6	0.930	0.0989
Ours (FFHQ)	**27.0**	**0.931**	**0.0975**

Table 3. **Novel view synthesis** – We sample pairs of images from one frame for each subject in the HUMBI dataset [73] and use them as query/target pairs. The query image is used to optimize a latent representation of the subject's face, which is then rendered from the target view. To evaluate how well the models have learned the 3D structure of faces, we then evaluate image reconstruction metrics for the face pixels of the predicted and target images after applying a mask computed from face landmarks.

4.3. Depth Prediction

We also evaluate our shape model by predicting depth values for images where ground truth depth is available. We

Method	Depth Correlation↑
Ground Truth	66
AIGN (supervised) [68]	50.81
DepthNetGAN (supervised) [45]	**58.68**
MOFA [65]	15.97
DepthNet [45]	35.77
UnSup3D [70]	54.65
Ours (CelebA-HQ)	50.18

Table 4. **Depth prediction** – Correlation between predicted and true keypoint depth values on 3DFAW. We compare to results from supervised and unsupervised methods as reported in [70] and [45].

use the 3DFAW dataset [21, 26, 27, 34, 75] which provides ground truth 3D keypoint locations. For this task we fit latent codes from our model on the 3DFAW images and sample the predicted depth values for each image-space landmark location. We use the same procedure as [70] to compute the correlation of the predicted and ground truth depth values, which is recorded in in Table 4. While our score is not as high as the best performing unsupervised method ([70]), it outperforms several supervised and unsupervised methods specifically designed for depth prediction.

4.4. High Resolution Image Synthesis

To demonstrate the benefits of being able to train directly on high-resolution images, we quantitatively and qualitatively compare high-resolution renders from LOLNeRF trained on 256×256 FFHQ and CelebA-HQ images to those of π-GAN [9] trained on 128×128 CelebA images (the largest feasible size used due to compute constraints). These results are shown in Table 2 and the supplementary material. We find that for this task our models do much better at reproducing high-resolution detail, even though both methods are capable of producing "infinite resolution" images in theory.

5. Discussion

5.1. Limitations and Future Work

While our method produces very high-quality results from training on in-the-wild data, it is still reliant on other methods for extracting semantic information (i.e. landmarks) about the objects being observed. As shown in Figure 9, this dependence can lead to failure cases for objects where the estimated pose or segmentation is incorrect. Finding a method to achieve alignment without any prior knowledge remains an open research question.

Also, while our auto-decoder framework has many advantages over GANs, it does not provide the same ability to maximize "plausibility" of the rendered images, which can result in some loss of detail. A possible direction for future work would be to augment our method with adversarial training to further improve the perceptual quality of images rendered from novel latent codes.

Figure 9. **Failure cases** – For some images, the segmentation module fails to remove foreground objects that are not the subject, or erroneously removes parts of the subject (Top). Our camera fitting also sometimes produces unreasonable camera estimates with focal lengths that are too short, or angles that are far from correct, resulting in distorted reconstruction of that subject during training (Bottom). Both failure modes occur in a small fraction our data, but the model still learns a space of 3D shape and appearance that performs well on structure-sensitive tasks, evident by our novel view synthesis (Table 3) and depth prediction (Table 4) results.

5.2. Ethical Considerations

Our research on image generation focuses on socially beneficial use cases and applications. When developed correctly, generative models can do that in a number of ways including simulating a diverse population of users (fairness) and amplify the effectiveness of personal data thus reducing the need for large scale data collection (privacy). However, we acknowledge the potential for misuse and importance of acting responsibly. To that end, we will only release the code for reproducibility purposes, but will not release any trained generative model.

5.3. Conclusions

We have presented LOLNeRF, a method for learning spaces of 3D shape and appearance from datasets of single-view images. Our experiments have shown that this method is able to learn effectively from unstructured, "in-the-wild" data, without incurring the high cost of a full-image discriminator, and while avoiding problems such as mode-dropping that are inherent to adversarial methods.

Acknowledgements

We thank Matthew Brown, Kevin Swersky, David Fleet, and Viral Carpenter for their helpful technical insights and feedback. Also Danica Matthews for her assistance with feline data collection. This work was partly supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Compute Canada.

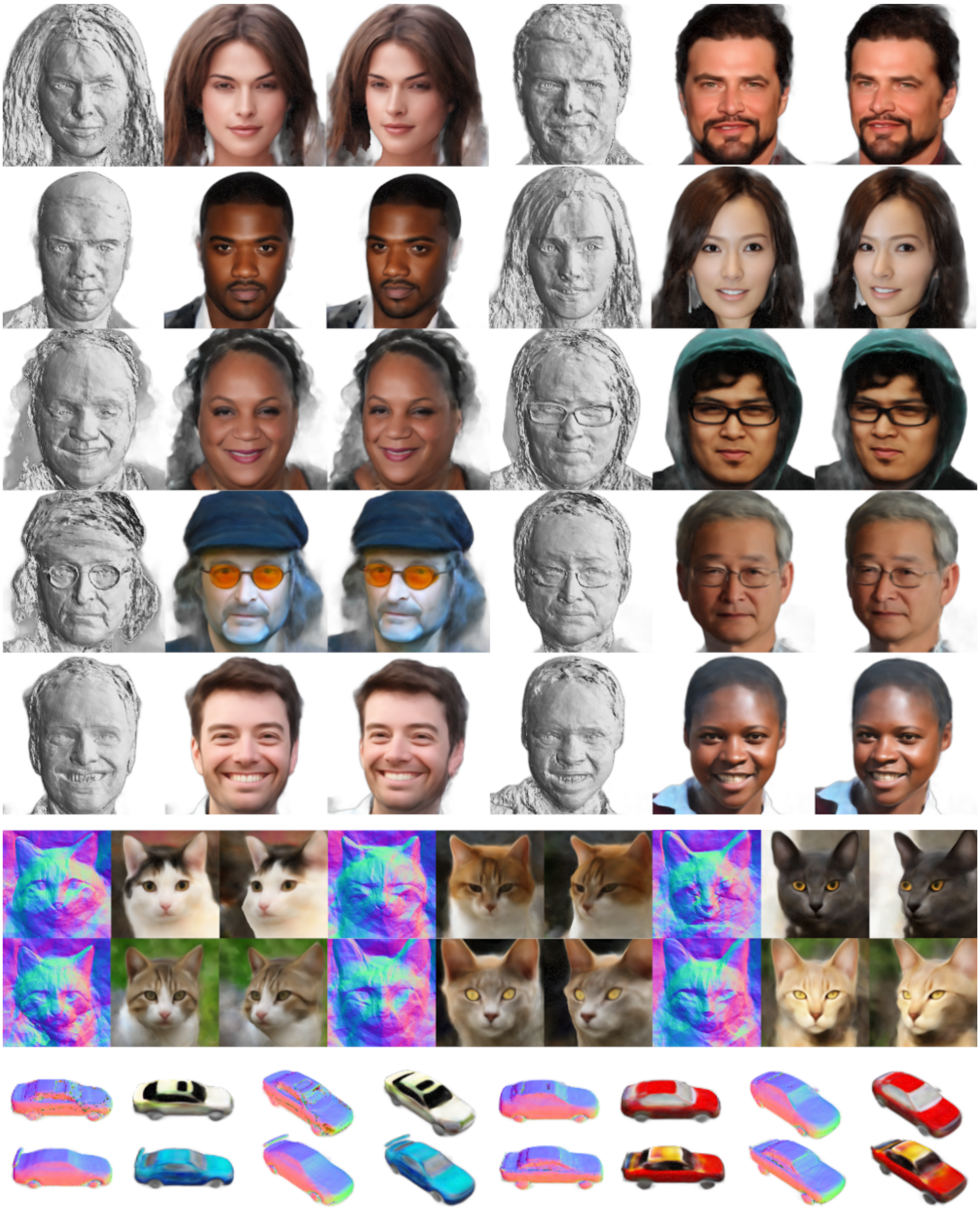

Figure 10. **Qualitative renders** – Outputs from our method on all datasets. Rows 1-2: CelebA-HQ [29], Rows 3-5: FFHQ [30], Rows 6-7: AFHQ Cats [12], Rows 8-9: SRN Cars [62].

References

[1] Alex M Andrew. Multiple view geometry in computer vision. *Kybernetes*, 2001. 1

[2] Idan Azuri and Daphna Weinshall. Generative latent implicit conditional optimization when learning from small sample. In *ICPR*, 2020. 3

[3] Jonathan Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *ICCV*, October 2021. 2

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 1

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Siggraph*, 1999. 2

[6] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the Latent Space of Generative Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 600–609. PMLR, July 2018. ISSN: 2640-3498. 3

[7] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. 1

[8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 5

[9] 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, 5, 6, 7

[10] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. BSP-Net: Generating Compact Meshes via Binary Space Partitioning. In *CVPR*, 2020. 1

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, pages 5939–5948, 2019. 2

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *CVPR*, pages 8188–8197, 2020. 5, 8

[13] Frank Dellaert and Yen-Chen Lin. Neural volume rendering: Nerf and beyond. *CoRR*, abs/2101.05204, 2021. 1, 2

[14] Boyang Deng, Kyle Genova, Soroosh Bouaziz, Geoffrey Hinton, Andrea Tagliasacchi, and Soroosh Yazdani. CvxNet: Learnable Convex Decomposition. In *CVPR*, 2020. 1, 3

[15] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2

[16] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *CVPR*, 2021. 1

[17] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic View Synthesis from Dynamic Monocular Video. In *ICCV*, October 2021. 4

[18] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William Freeman, and Thomas Funkhouser. Learning Shape Templates with Structured Implicit Functions. In *ICCV*, 2019. 1, 3

[19] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2

[20] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. In *Siggraph Asia*, 2018. 2

[21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 7

[22] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 2

[23] Philipp Henzler. Unsupervised Learning of 3D Object Categories from Videos in the Wild. In *CVPR*, 2021. 3

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[25] Wongbong Jang and Lourdes Agapito. CodeNeRF: Disentangled Neural Radiance Fields for Object Categories. In *ICCV*, October 2021. 1, 2

[26] László A. Jeni, Jeffrey F. Cohn, and Takeo Kanade. Dense 3D face alignment from 2D video for real-time use. *Image and Vision Computing*, 58:13–24, 2016. doi:10.1016/j.imavis.2016.05.009. 7

[27] László A. Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F. Cohn. The First 3D Face Alignment in the Wild (3DFAW) Challenge. In *ECCVW*, 2016. 7

[28] James T. Kajiya and Brian P. Von Herzen. Ray tracing volume densities. In *Siggraph*, 1984. 3

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, 2018. 5, 8

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 5, 8

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 2

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *NeurIPS*, 2018. 5

[33] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944. 1

[34] Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. A High-Resolution 3D Dynamic Facial Expression Database. In *The 8th International Conference on Automatic Face and Gesture Recognition (FGR08)*, Amsterdam, The Netherlands, 2008. 7

[35] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-Adjusting Neural Radiance Fields. In *ICCV*, October 2021. 2

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *CVPR*, 2015. 5

[37] Jitendra Malik, Pablo Arbeláez, João Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. The three R's of computer vision: Recognition, reconstruction and reorganization. *Pattern Recognition Letters*, 72:4–14, 2016. 1

[38] Marco Marchesi. Megapixel size image creation using generative adversarial networks. *arXiv preprint arXiv:1706.00082*, 2017. 5

[39] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 2, 3

[40] MediaPipe Face Mesh. `https://google.github.io/mediapipe/solutions/face_mesh.html`. Accessed: 2021-10-20. 4

[41] MediaPipe Selfie Segmentation. `https://google.github.io/mediapipe/solutions/selfie_segmentation.html`. Accessed: 2021-10-20. 4

[42] Quan Meng. GNeRF: GAN-Based Neural Radiance Field Without Posed Camera. In *ICCV*, October 2021. 2

[43] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*, 2019. 1, 2, 3

[44] 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, 3

[45] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised Depth Estimation, 3D Face Rotation and Replacement. In *NeurIPS*, 2018. 7

[46] 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. 1

[47] Michael Niemeyer and Andreas Geiger. CAMPARI: Camera-aware decomposed generative neural radiance fields. In *International Conference on 3D Vision (3DV)*, 2021. 1, 2, 3, 4

[48] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *CVPR*, 2021. 2

[49] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. In *ICCV*, October 2021. 2

[50] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[51] 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

[52] Keunhong Park, Utkarsh Sinha, Jonathan Barron, Sofien Bouaziz, Dan Goldman, Steven Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable Neural Radiance Fields. In *ICCV*, October 2021. 1, 2

[53] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 2021. 2

[54] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009. 2

[55] Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. In *Siggraph*, 2020. 2

[56] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016. 1

[57] 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. 1, 2, 3

[58] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20154–20166. Curran Associates, Inc., 2020. 1

[59] Sidak Pal Singh, Angela Fan, and Michael Auli. GLOSS: Generative latent optimization of sentence representations. *arXiv preprint arXiv:1907.06385*, 2019. 3

[60] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. In *NeurIPS*, 2020. 3

[61] Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021. 2

[62] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene Representation Networks: Continuous 3D-Structure-Aware Neural Scene Representations. In *Adv. Neural Inform. Process. Syst.*, 2019. 1, 2, 4, 5, 8

[63] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[64] Shufeng Tan and Michael L. Mayrovouniotis. Reducing data dimensionality through optimizing neural network inputs. *AIChE Journal*, 41(6):1471–1480, 1995. 3

[65] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 7

[66] Products - 3dMD. `https://3dmd.com/products/`. Accessed: 2021-10-25. 2

[67] Alex Trevithick and Bo Yang. GRF: Learning a General Radiance Field for 3D Scene Representation and Rendering. In *ICCV*, October 2021. 3

[68] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372. IEEE, 2017. 7

[69] 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

[70] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 2, 3, 7

[71] Christopher Xie, Keunhong Park, Ricardo Martin-Brualla, and Matthew Brown. FiG-NeRF: Figure-ground neural radiance fields for 3d object category modelling. *arXiv preprint arXiv:2104.08418*, 2021. 2

[72] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*, 2021. 3

[73] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *CVPR*, 2020. 6

[74] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NERF++: Analyzing and improving neural radiance fields. *https://arxiv.org/abs/2010.07492*, 2020. 2, 4

[75] Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. A high-resolution spontaneous 3D dynamic facial expression database. *Image and Vision Computing (special issue of The Best of Face and Gesture 2013)*, 32:692–706, 2014. 7