LOLNeRF: Learn from One Look

Supplementary Material

A. Qualitative Results – Figures 11, 12

We provide additional qualitative comparisons with π-GAN [9] in Figures 11 and 12, which show novel views at high angles and high resolution renders, respectively.

Please also see the project website (https://lolnerf.github.io) for animations which demonstrate the high-quality 3D structure and novel views that our method produces.

B. Camera Fitting Procedure

For a class-specific landmarker which provides estimates for M 2D landmarks $\ell \in \mathbb{R}^{M\times 2}$, we estimate the extrinsics $\mathbf{T}$ and (optionally) intrinsics $\mathbf{K}$ of a camera which minimizes the reprojection error between ℓ and a set of canonical 3D positions $\mathbf{p} \in \mathbb{R}^{M\times 3}$. We achieve this by solving the following least-squares optimization:

$$\arg\min_{\mathbf{T},\mathbf{K}} ||\ell - P(\mathbf{p}|\mathbf{T},\mathbf{K})||^2 \tag{10}$$

where $P(\mathbf{x}|\mathbf{T},\mathbf{K})$ represents the projection operation for a world-space position vector $\mathbf{x}$ into image space. We perform this optimization using the Levenberg–Marquardt algorithm [33]. The canonical positions $\mathbf{p}$ may be either manually specified or derived from data. For human faces we use a predetermined set of positions which correspond to the known average geometry of the human face. For AFHQ, we perform a version of the above optimization jointly across all images where $\mathbf{p}$ is also a free variable, and constrained only to obey symmetry.

In our experiments we predict camera intrinsics for human face data, but use fixed intrinsics for AFHQ where the landmarks are less effective in constraining the focal length. For SRN cars, we use the camera intrinsics and extrinsics provided with the dataset, though we note that semantically consistent landmarkers do exist for this class of data [63].

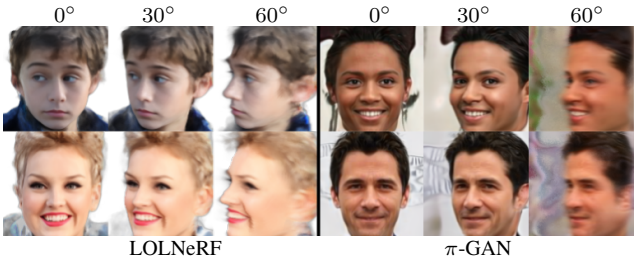


Figure 11. **High view angles** – both methods show degraded quality at high angles, but ours maintains better sharpness and view-consistency.

Method	FID↓	KID↓	IS↑
HoloGAN [46]	39.7	2.91	1.89
GRAF [58]	41.1	2.29	2.34
π-GAN [9]	**14.7**	0.39	**2.62**
Ours	128.2	**0.11**	2.34

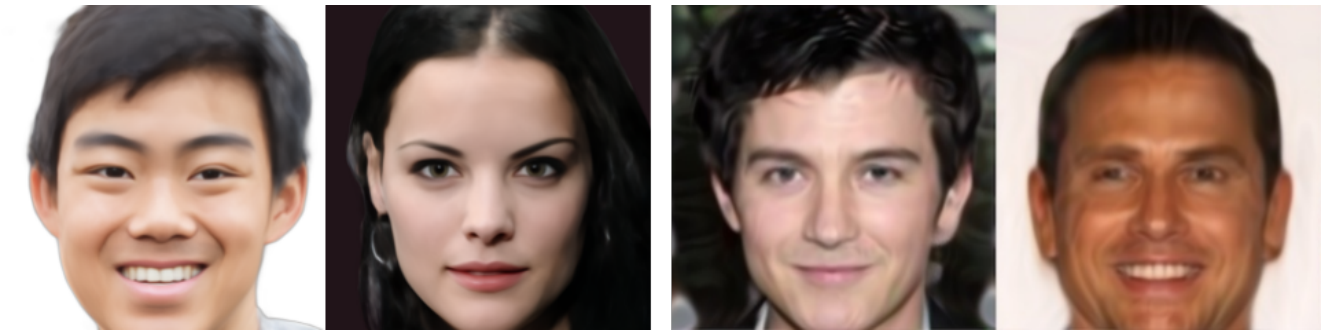
Table 5. **Unconditional Sampling Quality** – Perceptual image distribution quality metrics on CelebA for our model and baselines. The results for HoloGAN and GRAF are taken from [9].

C. Dataset Size Ablation – Figure 13

To quantify our method's dependence on large amounts of data, we perform an ablation study in which we train models with subsets of the full dataset. We then show the novel view synthesis quality from these networks as a way of determining how well they have generalized to reconstructing different views. The results of this experiment are shown in Figure 13. We find a clear trade-off in quality of the training image reconstruction and quality of the learned 3D structure as the dataset size increases. Very small datasets reconstruct their training images with high accuracy, but produce completely unreasonable geometry and novel views. As the number of training images increases, the accuracy of reconstruction slowly decreases, but the predicted structure generalizes to become much more consistent and geometrically reasonable.

D. Unconditional Sampling – Figure 14

To evaluate the quality of unconditional samples generated by our PCA-based sampling method, we compute three standard quality metrics for generative image models on these renders: Frechet Inception Distace (FID) [24], Kernel Inception Distance, (KID) [4], and Inception Score (IS) [56], the results of which are shown in Table 5. We find that our method achieves an inception score competitive with other 3D-aware GAN methods, indicating that we are able to model a variety of facial appearances. Our result for the distribution distance metrics, FID and KID, however, show opposing results with our method doing far worse in FID but better in KID. The reason for this is not entirely clear, but FID has been shown to be sensitive to noise [7], and details in the peripheral areas of our generated images show more noise-like artifacts than π-GAN [9]. Regardless, we do not necessarily expect to outperform π-GAN [9] for these metrics as it is trained to produce images with high perceptual quality as determined by a CNN, which is a better proxy for these distribution metrics than our reconstruction loss.



Our Method π-GAN [9]

Figure 12. **High resolution renders** – Our method is trained natively to reconstruct high resolution images and can reconstruct sharp details, while π-GAN [9] reconstructions lack detail beyond its training resolution and contain artifacts that become easily visible.

10 50 100 500 1k 5k 10k

Figure 13. **Dataset size ablation** – we show the behaviour of the method as the size of the training dataset varies. Rows 1, 3, and 5 show the learned reconstruction of training images from the predicted view for that image. Rows 2, 4, and 6 show the models rendered from a novel view. The columns show the results as the total number of training images is increased from ten to ten thousand. Note that this ablation is performed with lower-capacity networks (256 neurons per layer), both to avoid wasting energy in training and to more clearly show how reconstruction quality changes with dataset size.

E. Architecture Details

Our architecture uses a standard NeRF backbone architecture as described in [44] with a few modifications. In addition to the standard positional encoding we condition the network on an additional latent code by concatenating it alongside the positional encoding. For SRN cars and AFHQ we use the standard 256 neuron network width and 256-dimensional latents for this network, but we increase to 1024 neurons and 2048-dimensional latents for our high-resolution CelebA-HQ and FFHQ models. For our background model we use a 5-layer, 256-neuron relu MLP in all cases. During training, we use 128 samples per ray for volume rendering with no hierarchical sampling.

Our Method π-GAN [9]

Figure 14. **Unconditional generation** – Both methods produce samples that resemble the training distribution in appearance and shape. Ours produce sharper details in the central face region where the data is more consistent, while π-GAN [9] produces more plausible hair and backgrounds due to its adversarial training.

F. Training Details

We train each model for 500k iterations using a batch size of 32 pixels per image, with a total of 4096 images included in each batch. For comparison, with 256^2 images, this compute budget would allow just 2 images per batch for a GAN-based method which renders entire frames.

We train with an ADAM [31] optimizer using exponential decay for the learning rate from 5×10^{-4} to 1×10^{-4}. We run each training job using 64 v4 Tensor Processing Unit chips, taking approximately 36 hours to complete for our high resolution models.