

Pixel-aligned Volumetric Avatars

– Supplemental Document –

Amit Raj¹ Michael Zollhöfer² Tomas Simon² Jason Saragih²
Shunsuke Saito² James Hays¹ Stephen Lombardi²

¹ Georgia Institute of Technology ² Facebook Reality Labs Research

1. Architecture Details

1.1. Feature Extraction Network

Fig. 2 shows the architecture of the feature extraction network. We employ a shallow convolutional encoder-decoder network with no narrow bottlenecks to retain as much local information as possible. This is in contrast to U-Net or Hour Glass feature extraction networks. We also present the cNeRF baseline architecture used for comparison, see Fig. 4

1.2. Positional Features

To generate positional features, we use positional encoding as described in Mildenhall et al [1]. Particularly, for each point along the ray \mathbf{X} and view direction \mathbf{d} , we calculate the encoding as follows:

$$\phi(X) = [\sin(2^0 \pi X), \cos(2^0 \pi X), \dots, \sin(2^l \pi X), \cos(2^l \pi X)]$$

Where we choose $l = 9$ for encoding the position and $l = 3$ for encoding the direction.

1.3. Radiance Field Network

Fig. 3 shows the architecture of our radiance field network. Our architecture is inspired by Mildenhall et al.[1], but we employ more view-dependent layers at the end of the network.

2. Additional Results

2.1. Hair integration

One advantage of volumetric models is that models learned from different sources can be combined seamlessly, since the rendering involves a ray marching step through the learned volume. Fig 5 shows the integration of an independently learned hair volume with a radiance field learned using pixel aligned volumetric avatars.

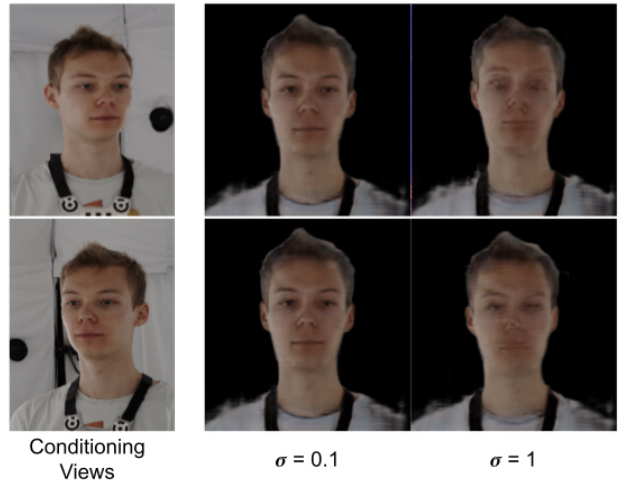


Figure 1. Robustness test. Pose with added Gaussian noise of standard deviation of 0.1 and 1.0 (top: noise added to camera translation, bottom: noise added to camera rotation).

2.2. View Synthesis

Fig. 7 and Fig. 8 demonstrate additional view synthesis results on novel identities.

2.3. Sensitivity to camera parameters

We test the sensitivity of our approach to noise in the pose estimate. Our models, similar to NeRF, are sensitive to errors in camera pose during training. At test time, it is more robust due to information averaging from multiple views, but shows ghosting artifacts under large noise applied to translation or rotation parameters, as shown in Fig. 1.

2.4. Failure case

Fig. 6 shows a failure case of capturing the geometry of glasses for *out-of-distribution* data.

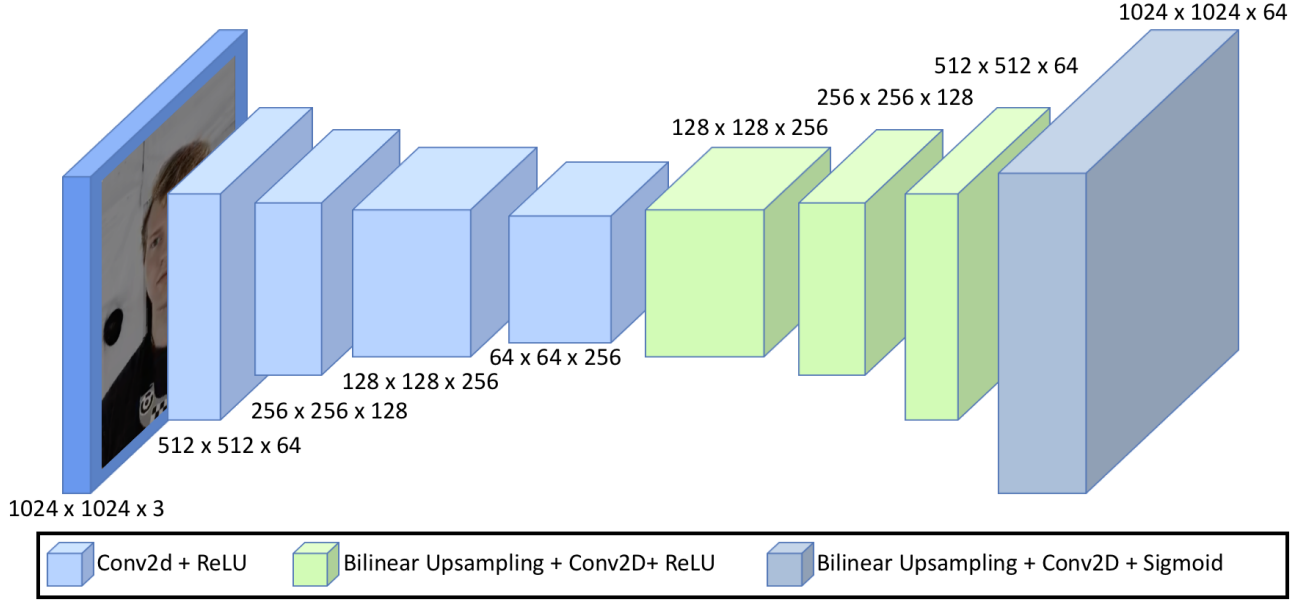


Figure 2. Architecture of the feature extraction network. We use a shallow convolutional encoder-decoder network to retain local information to a greater extent. The output feature map has the same spatial dimensions as the input image.

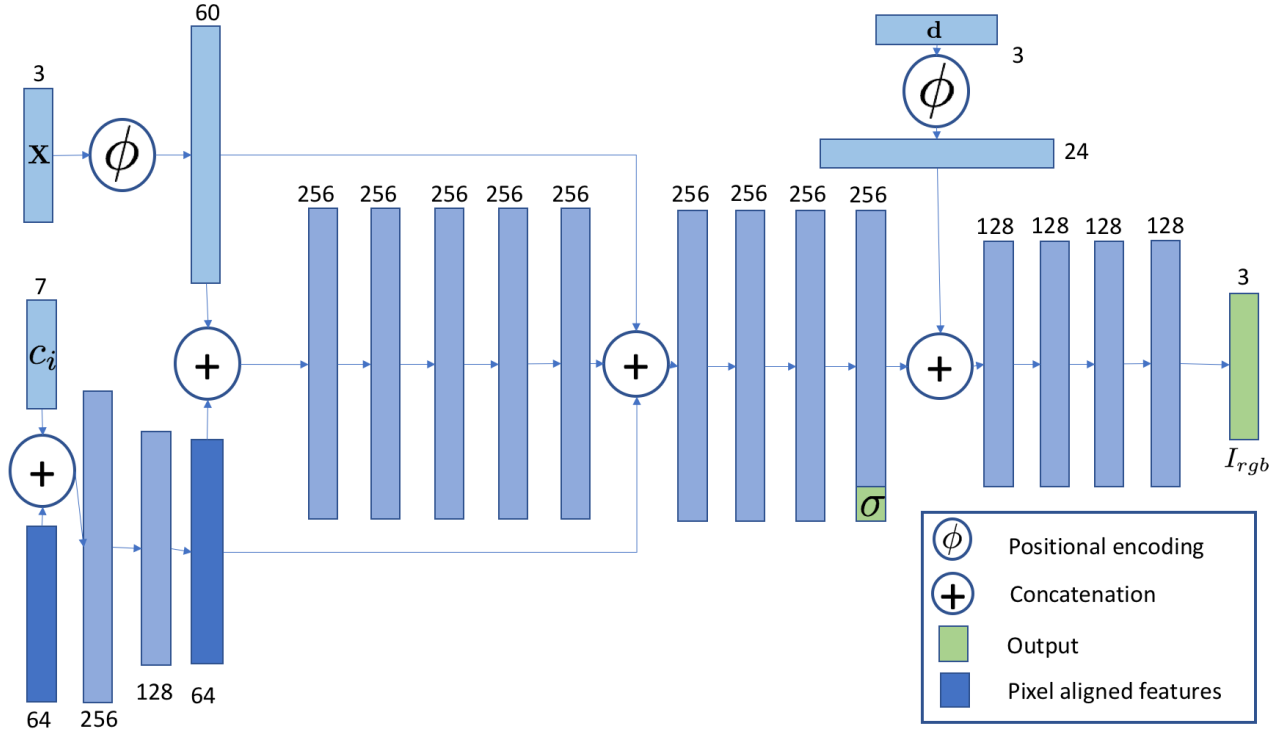


Figure 3. Architecture of our radiance field network. We use the camera information c_i to generate camera summarized features to allow for feature aggregation from different conditioning viewpoints.

References

- [1] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020. 1

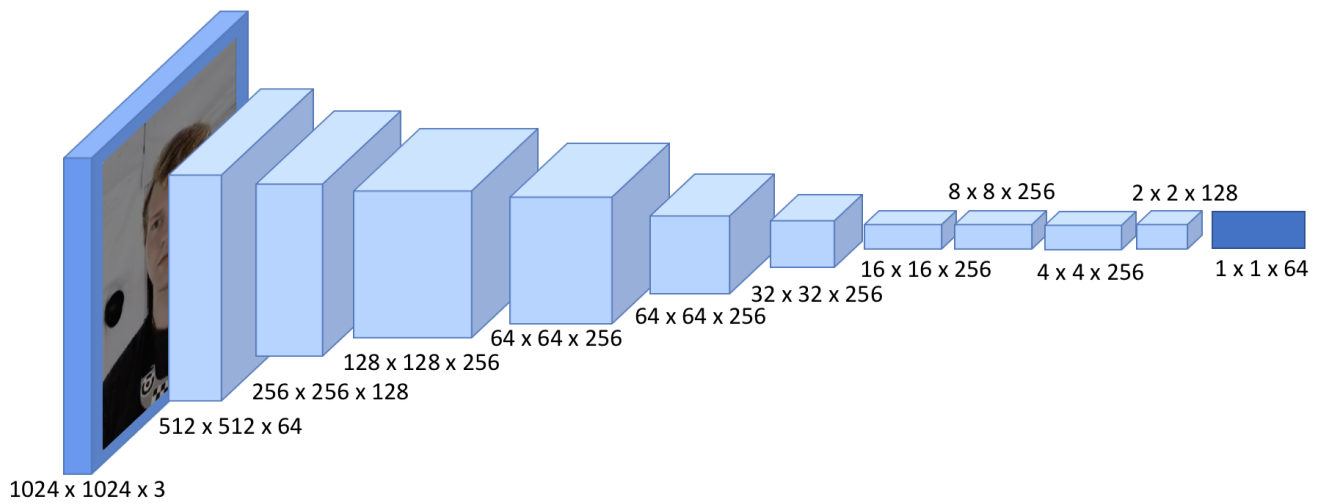


Figure 4. Architecture of the cNerf baseline. This baseline learns a global encoding for each conditioning viewpoint and retains no local pixel level information that can be used by the radiance field network.



Figure 5. Independently learned hair volume integrated with a pixel-aligned volumetric avatar. The hair volume is learned using pre-computed hair segmentation supervision.

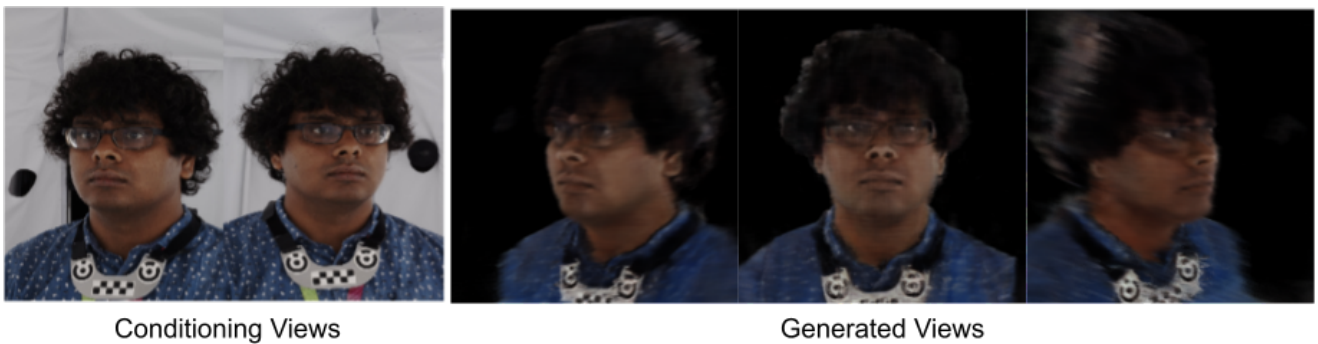


Figure 6. Generalization to out-of-distribution data (left: inputs, right: results) fails to capture geometry of glasses

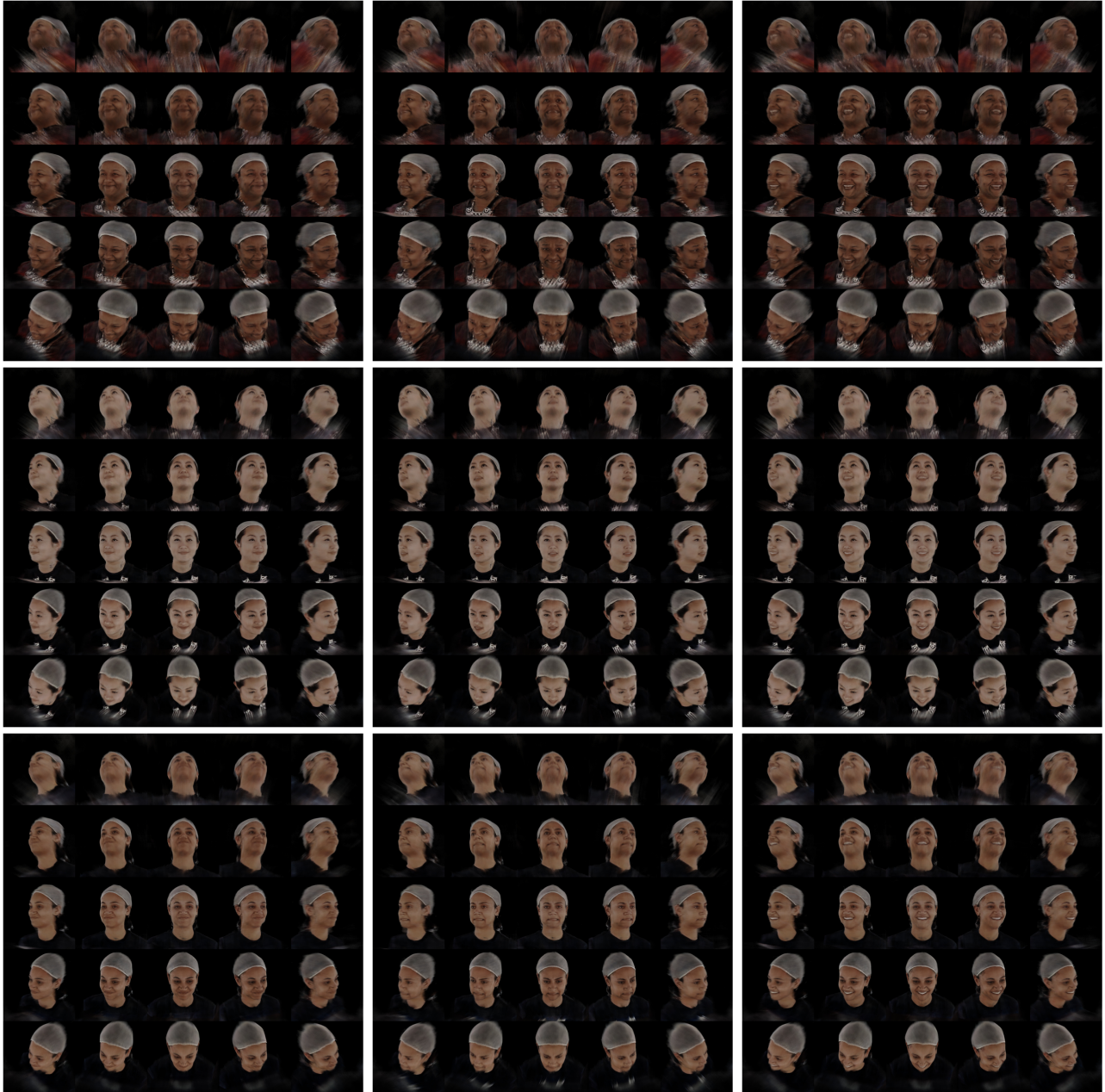


Figure 7. Additional novel view synthesis results. Here we show 3 novel identities in 3 different expressions from 25 views not seen during training. Our method is able to retain high-frequency detail, producing high-quality avatars from multiple views.

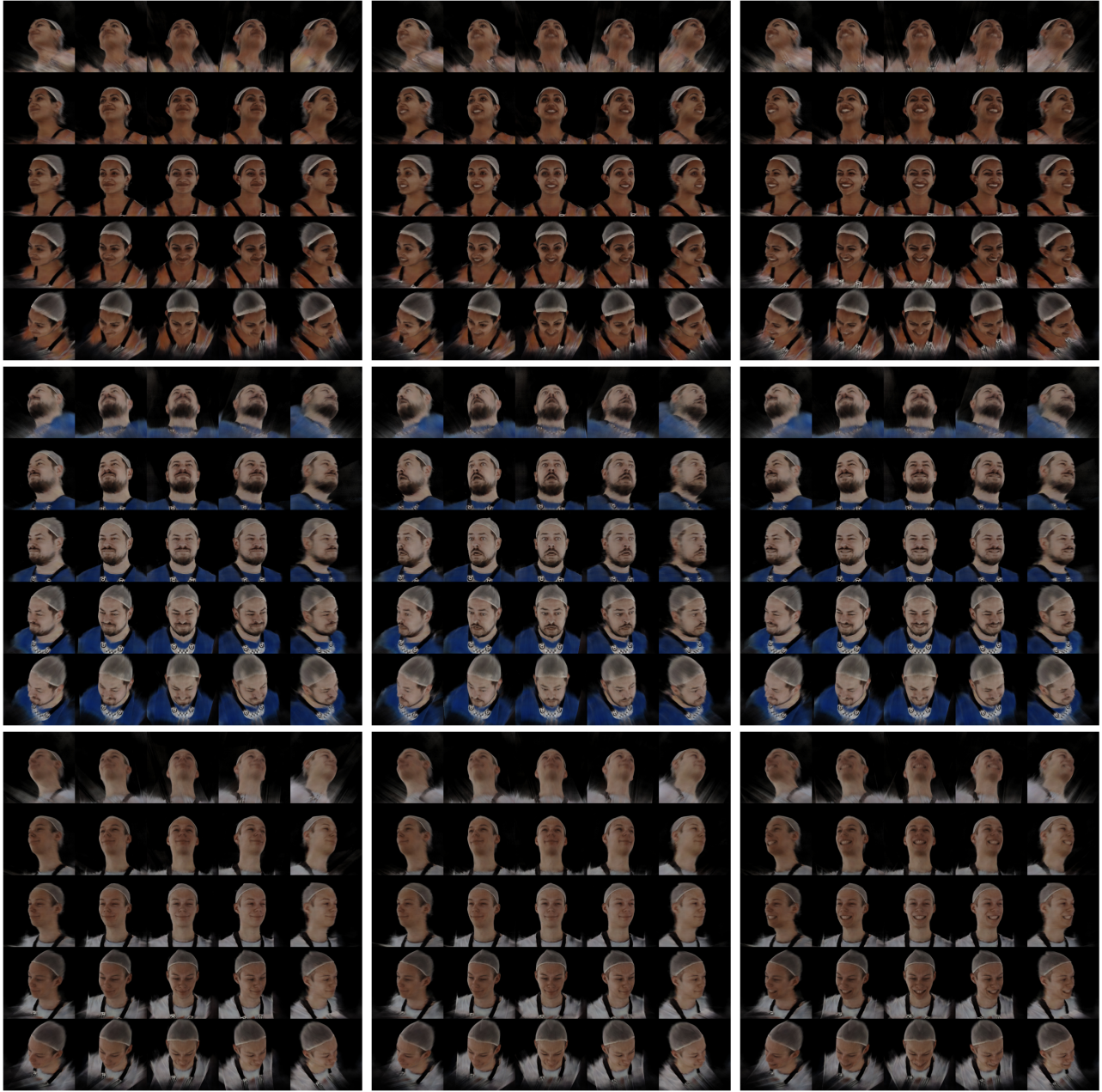


Figure 8. Additional novel view synthesis results. Here we show 3 novel identities in 3 different expressions from 25 views not seen during training. Our method is able to retain high-frequency detail, producing high-quality avatars from multiple views.