

– Supplemental Document –

Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction

Guy Gafni¹ Justus Thies¹ Michael Zollhöfer² Matthias Nießner¹
¹Technical University of Munich ²Facebook Reality Labs Research

1. Network Architecture

We provide additional details of the proposed dynamic neural radiance fields architecture. As mentioned in the main paper, the dynamic neural radiance field is represented as a multi-layer perceptron (MLP). In Fig. 1, we depict the underlying architecture.

The dynamic neural radiance field is controlled by the expression coefficients that correspond to the blendshape basis of the used face tracker [2]. To compensate for missing information, we also feed in the learned latent codes γ . For a given sample location (x, y, z) and the corresponding viewing direction \vec{d} , the MLP outputs the color and density which is used for the volumetric rendering, explained in the main document.

The MLP is based on a backbone of 8 fully-connected layers, each 256 neurons-wide, followed by ReLu as activation functions. These activations are fed through a single layer to predict the density value, as well as a 4-layer, 128 neuron-wide branch to predict the final color value of the query point.

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. In *ECCV*, 2020. 2
- [2] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016. 1

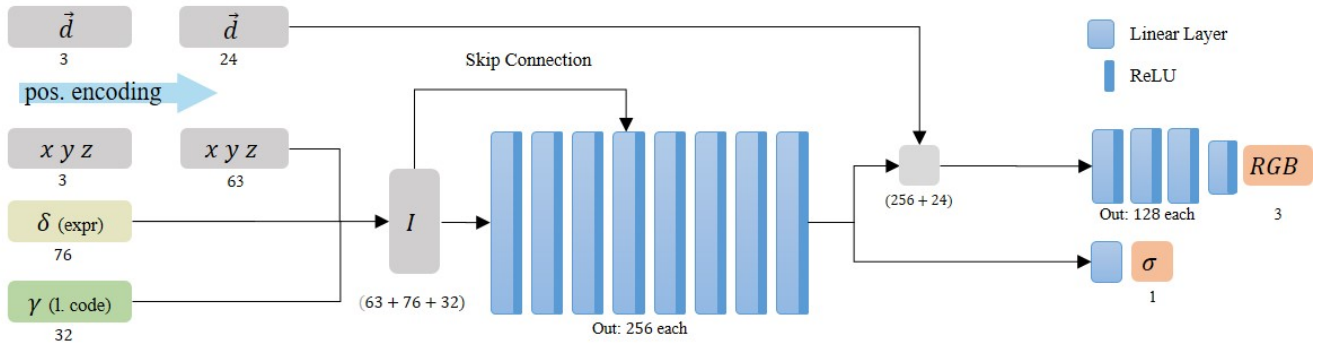


Figure 1: Our Dynamic Neural Radiance Field is represented as a multi-layer perceptron (MLP). As input it gets the viewing direction \vec{d} , the sample position (x, y, z) , the expression coefficients δ as well as the learned latent codes γ . The viewing direction as well as the position are encoded using positional encoding [1]. The MLP consists of a backbone with 8 linear layers each with ReLU non-linearity which takes the position, the expression and latent code as input (concatenated as vector I). The output of the backbone is used to compute the density σ . To compute the color, the output of the backbone is concatenated with the encoded viewing direction and inputted into another 4 linear layers with ReLU activations.