Space-time Neural Irradiance Fields for Free-Viewpoint Video Supplementary Material

Wenqi Xian* Cornell Tech

Jia-Bin Huang Virginia Tech

Johannes Kopf Facebook

Changil Kim Facebook

1. Overview

This supplementary document provides more implementation details (in Section 2), training details (in Section 3) and additional quantitative comparison in Section 4. We include additional qualitative results in our project website https://video-nerf.github.io. In particular, we test on 10 different videos and provide comparison with baselines and different loss configurations.

2. Implementation Details

We implement our framework using PyTorch. In all our experiments, we empirically set the hyper-parameters as $\alpha = 1$, $\beta = 100$, and $\gamma = 10$.

We calculate all the losses except the static scene loss on a batch of $N_\mathrm{r} = 1024$ rays that are randomly drawn from an input frame I_t without replacement. We randomly choose $N_\mathrm{s} = 1024$ from $\mathcal{X}$ at each step (also without replacement) for the static scene loss. We normalize the time t such that $\mathcal{T} = [-1, 1]$ and apply the positional encoding with 4 frequency bands. Following NeRF [1], we apply positional encoding to spatial positions $\mathbf{x}$. While we do not use the normalized device coordinates, we sample each ray uniformly in *inverse* depth. We set the depth range z_n and z_f as the global minimum and maximum of all frames' depth values.

3. Training Details

We used the same MLP architecture as in [1], except that we use 1024 activations for the first 8 MLP layers instead of 256. Our models are trained with various combinations of the losses presented in the main paper but otherwise with the same hyperparameters. We used the Adam optimizer with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a learning rate of 0.0005. We train the MLP for 800k iterations. It takes about 48 hours to train a network with about 100 video frames at the 960×540 resolution with 4 NVIDIA V100 GPUs.

4. Additional comparison to baseline

In this section, we provide a quantitative comparison to the inpainted mesh method. We rendered the ground-truth mesh provided by the Sintel dataset in 2D and inpaint the missing pixels in disoccluded areas using state-of-the-arts video completion algorithms. Then, we evaluate how well our method handles disoccluded areas, compared with the inpainted mesh method, for which we used the same Sintel GT depth. Table 1 shows the quantitative evaluation measured in PSNR metric. It demonstrates that inpainting is not sufficient to get good disocclusion contents, and validates that our approach produces significantly fewer artifacts than the baseline method using video completion.

Table 1. Reported PSNR on disoccluded pixels only.

Methods	Bandage1	Bandage2	Sleeping1	Sleeping2	Alley1	Bamboo1
Inpainted	**16.52**	13.32	20.08	17.52	14.82	12.23
Ours	15.54	**21.99**	**21.65**	**33.91**	**26.75**	**25.28**

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. 1