Neural 3D Video Synthesis from Multi-view Video

Tianye Li[1,3,*] Mira Slavcheva[1,*] Michael Zollhoefer[1] Simon Green[1] Christoph Lassner[1]
Changil Kim[2] Tanner Schmidt[1] Steven Lovegrove[1] Michael Goesele[1] Richard Newcombe[1] Zhaoyang Lv[1]

[1] Reality Labs Research [2] Meta [3] USC

* equal contributions

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Background

Goal: 3D video synthesis

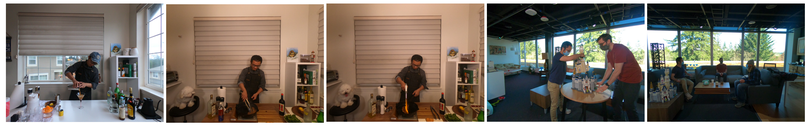
Capture dynamic scenes

- Non-trivial to extend NeRF to dynamic case
- Non-rigid motion, volumetric and topology changes

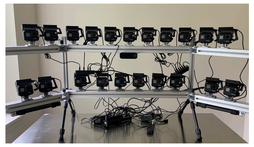
Long training time

- NeRF: 50 GPU hours, 10 sec, 30fps = 15,000 GPUS hours

Data



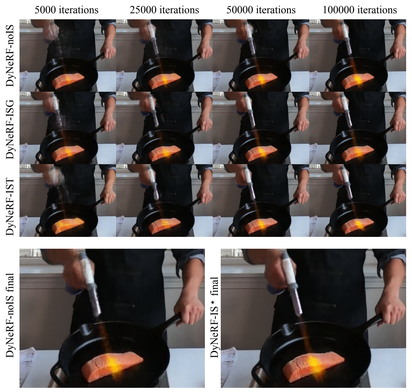
- The datasets cover challenging dynamic objects and view-dependent effects in a natural daily indoor environment
- We release the datasets for research purposes at our project page



Multi-camera rig

An example of the training views (right) and test view (left)

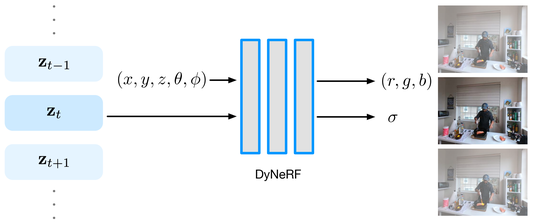
Results



Our 3D video results on the Broxton et al. datasets with a different capture setting. Please note that our representation is compact (28MB for 150 frames).

(left) The efficient training strategies accelerate training and improves visual quality.

DyNeRF: Dynamic Neural Radiance Field

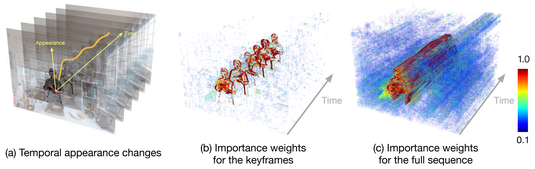


High-dim. latent codes to capture scene motion and dynamic appearances

Compactness: 10 seconds 30 FPS 2.7K resolution, 18-view videos can be compressed to only 28 MB

Space-time continuity: synthesis from arbitrary views and time

Efficient Training Strategies



(a) Temporal appearance changes

(b) Importance weights for the keyframes

(c) Importance weights for the full sequence

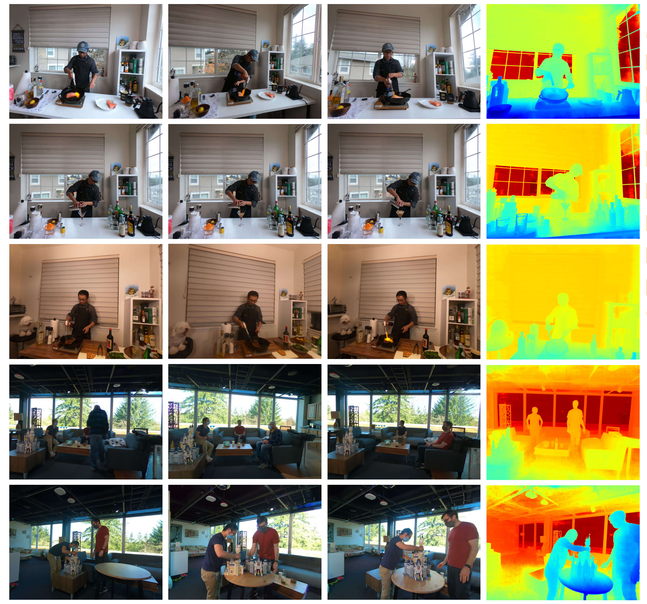
$$\mathcal{L}_{\text{efficient}} = \sum_{t \in \mathcal{S}, \mathbf{r} \in \mathcal{I}} \sum_{j \in \{c,f\}} \left\| \hat{\mathbf{C}}_j^{(t)}(\mathbf{r}) - \mathbf{C}^{(t)}(\mathbf{r}) \right\|_2^2$$

Hierarchical training
- First train on keyframes
- Then optimize for full sequences.

Ray importance sampling
- Explore spatial-temporal redundancy
- Emphasize on highly time-variant rays (pixels)

Time (frame) selection
Space (ray) selection
Stage (coarse, fine) as in NeRF

Results (cont.)



Dataset, Results, More Info

https://neural-3d-video.github.io/

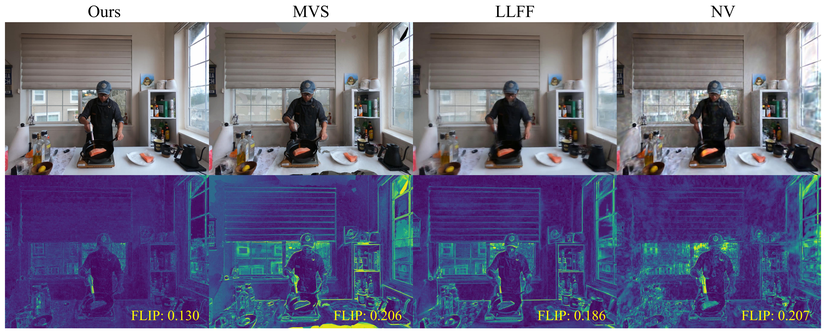
Our method synthesizes high-quality 3D videos for various dynamic real-world scenes.

By distilling the pre-trained DyNeRF model into layers meshes, we can render interactive 3D videos on a Quest 2 VR headset.

Method	PSNR ↑	MSE ↓	DSSIM ↓	LPIPS ↓	FLIP ↓
MVS	19.1213	0.01226	0.1116	0.2599	0.2542
NeuralVolumes	22.7975	0.00525	0.0618	0.2951	0.2049
LLFF	23.2388	0.00475	0.0762	0.2346	0.1867
NeRF-T	28.4487	0.00144	0.0228	0.1000	0.1415
DyNeRF†	28.4994	0.00143	0.0231	0.0985	0.1455
DyNeRF	**29.5808**	**0.00110**	**0.0197**	**0.0832**	**0.1347**

Our method outperforms existing methods and baseline methods in all visual quality metrics.

Results (cont.)



Our method achieves the best visual quality compared to the existing methods.

Limitation and Future Work

Challenge: outdoor scenes
- changing illuminations between cameras
- larger scene volume with complex geometries

References

- Mildenhall et al., NeRF: Representing scenes as neural radiance fields for view synthesis, ECCV 2020
- Mildenhall et al., Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, SIGGRAPH 2019
- Lombardi et al., Neural volumes: learning dynamic renderable volumes from images, SIGGRAPH 2019
- Broxton et al., Immersive light field video with a layered mesh representation, SIGGRAPH 2020