TeTriRF: Temporal Tri-Plane Radiance Fields for Efficient Free-Viewpoint Video

Supplementary Material

1. Implementation Details

Model Configuration. In our setup, all frames within a sequence utilize a common bounding box that defines their world space. These bounding boxes are derived based on the camera configurations. For object-centric datasets (NHR and ReRF), the world size is set to 120^3, while for the DyNeRF dataset, it is 210^3. We determine the feature plane resolution as three times the world size. Specifically, this results in approximately 360×360 for the NHR and ReRF datasets and 600×600 for the DyNeRF dataset, aiming to capture high-frequency signals effectively. Each feature plane comprises $h = 10$ channels, leading to a concatenated feature vector for each 3D point with a dimensionality of 30. The viewing directions undergo positional encoding with 4 frequency levels. We combine these encoded viewing directions with point feature vectors to serve as inputs for the MLP decoder Φ. The decoder Φ is a three-layer multilayer perceptron, having a width of 128. It outputs the RGB value for the sampled point. A unique MLP decoder is allocated to each frame group, facilitating shared learning across the frames in a group.

Training. For training, we employ the Adam optimizer [4] to update the density grids, tri-planes, and MLP decoder weights. The respective learning rates for these components are set to $1.5e^{-1}$ for the density grids and tri-planes, and $1e^{-3}$ for the MLP decoder. We implement group-based regularization with weights $\lambda_1 = 1e^{-3}$ and $\lambda_2 = 2e^{-3}$. Each training batch processes 17800 sampled rays from the dataset, and we conduct 40000 training iterations for each group. In our progressive scaling approach, the hybrid representation is upscaled by a factor of two at specific iterations: $[1000, 2000, 3000, 4000]$ during the first pass, and $[9000, 11000, 13000]$ during the second. We downscale the resolutions at the first and 7000-th iterations. Full resolutions are achieved at the 4000-th and 13000-th iterations. Every 1000 iterations, we update the occupancy grids V_o based on the density grids V_σ, fomulated as

$$V_o = \rho(\kappa(1 - \frac{1}{1 + \exp(V_\sigma)}), \lambda_{th}). \quad (1)$$

Here, $\kappa(\cdot)$ represents a 3D max pooling function with a 3×3 kernel, and $\rho(\cdot)$ is a thresholding function that outputs 1 if the grid element is greater than $\lambda_{th} = 1e^{-4}$, otherwise 0, indicating occupancy. At the 13000-th iteration, we filter out rays that do not intersect with any objects according to the current occupancy grid.

		PSNR	SSIM	LPIPS	Size (KB)
sport1	K-plane	30.40	0.962	0.0615	2986
	HumanRF	32.39	0.885	0.0318	2852
	TiNeuVox	30.54	0.961	0.0831	5580
	ReRF	30.83	0.973	0.0505	1113
	Ours(low)	31.79	0.969	0.0516	11.01
	Ours(high)	33.41	0.980	0.0389	79.92
sport2	K-plane	32.10	0.975	0.0472	2986
	HumanRF	33.04	0.889	0.0316	2852
	TiNeuVox	32.97	0.972	0.0568	5580
	ReRF	31.83	0.976	0.0487	1316
	Ours(low)	31.75	0.973	0.0498	10.56
	Ours(high)	34.14	0.983	0.0383	75.85
sport3	K-plane	30.20	0.962	0.0610	2986
	HumanRF	32.11	0.885	0.0328	2852
	TiNeuVox	30.11	0.960	0.0696	5580
	ReRF	30.89	0.976	0.0473	1243
	Ours(low)	30.38	0.967	0.0546	12.96
	Ours(high)	32.90	0.980	0.0394	94.58
basketball	K-plane	28.02	0.957	0.0822	2986
	HumanRF	30.09	0.829	0.0469	2852
	TiNeuVox	28.18	0.956	0.0991	5580
	ReRF	27.82	0.963	0.0747	1208
	Ours(low)	27.79	0.957	0.0806	12.53
	Ours(high)	29.85	0.970	0.0649	90.97

Table 1. Per-scene results on NHR dataset [11]. Values are averaged out over the number of frames in each scene.

2. Comparison Setups

We compared TeTriRF with several contemporary dynamic NeRF techniques, including KPlanes [2], HumanRF [3], TiNeuVox [1], and ReRF [10]. For forward-facing scenes, TeTriRF was additionally compared with NeRFPlayer [6].

KPlanes. We use the official implementation from NeRFStudio [7]. The KPlanes model was jointly trained on the entire sequence for 50,000 iterations, using a grid size of 256^3 and a time resolution of 100, as recommended.

HumanRF. We employed their official code for our experiments. Two hundred frames were trained jointly over 50,000 iterations. Initially, occupancy grids were generated using foreground masks as outlined in [3], followed by their prescribed training steps.

TiNeuVox. We use their official code. Due to memory constraints, each sequence was split into eight groups of 25 frames each. We used a grid size of 180^3 and trained each

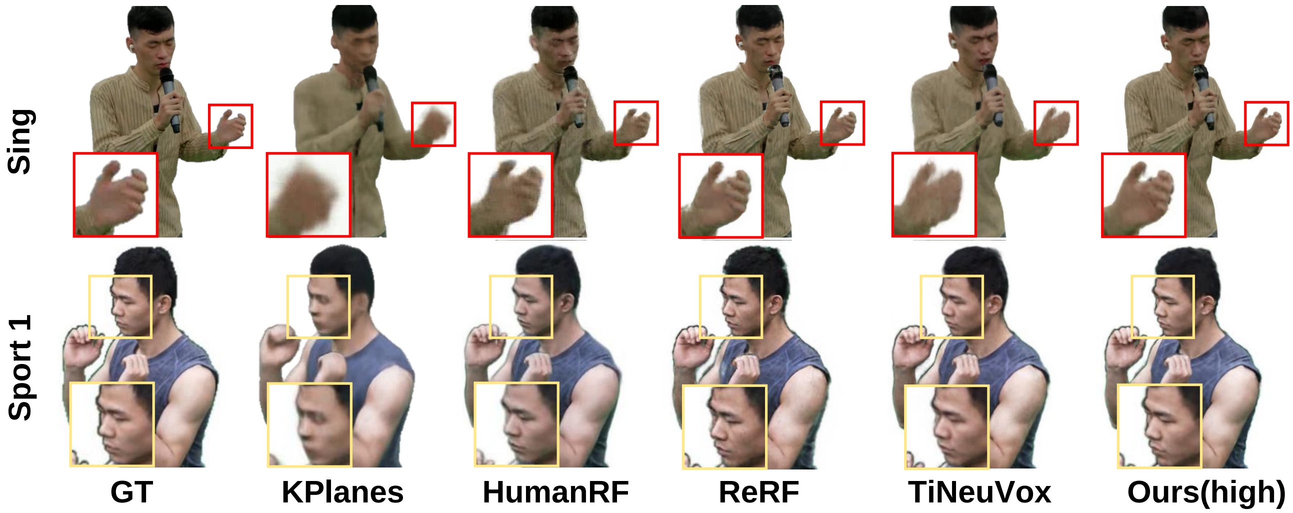

Figure 1. Extra qualitative results on 'Sing' from ReRF dataset [10] and 'Sport1' from NHR dataset [11].

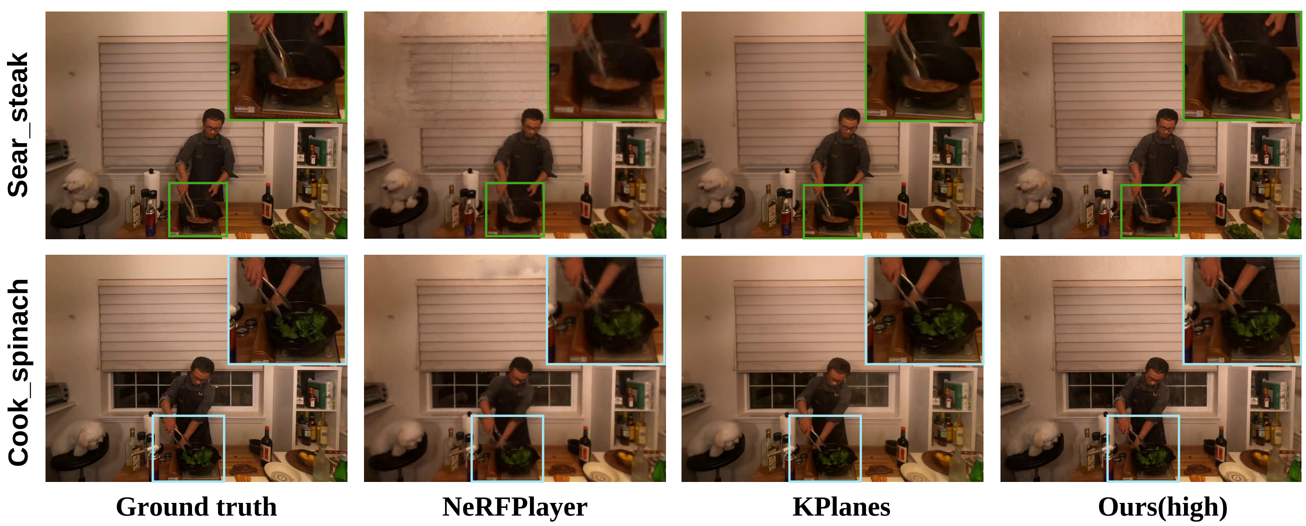

Figure 2. Extra qualitative results on DyNeRF dataset [5].

group for 30,000 iterations.

ReRF. The official code and default settings were used in our experiments, compressing sequences with a quality factor of 99.

NeRFPlayer. We relied on the quantitative results reported in the original paper [6] and conducted qualitative analyses using the official NeRFStudio implementation based on Nerfacto under default settings.

MixVoxels. We conducted tests using its official pretrained model, MixVoxels-M [8], under identical experimental settings and present only the averaged results.

Ours. Following the configurations detailed in our implementation section, we leverage the FFMPEG software with the libx265 codec for compressing the feature and density image sequences.

3. More Results

Table 1, Table 2, and Table 3 provide the detailed results for each scene. Figure 1 and Figure 2 demonstrate the qualitative comparison on three datasets.

4. Video

Please refer to our project page(`https://wuminye.github.io/projects/TeTriRF/`) for more qualitative results and comparisons. We use 'Ours(high)' in our video.

References

[1] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian.

		PSNR	SSIM	LPIPS	Size (KB)
box	K-plane	27.96	0.952	0.0836	2986
	HumanRF	29.07	0.884	0.0614	2852
	TiNeuVox	31.11	0.962	0.0633	5580
	ReRF	30.97	0.972	0.0516	925
	Ours(low)	27.94	0.955	0.0655	11.86
	Ours(high)	31.39	0.968	0.0498	70.01
kpop	K-plane	26.95	0.954	0.0984	2986
	HumanRF	28.84	0.901	0.0682	2852
	TiNeuVox	27.22	0.952	0.0887	5580
	ReRF	31.94	0.976	0.0436	725
	Ours(low)	27.03	0.964	0.0678	13.11
	Ours(high)	30.25	0.977	0.0526	80.1
sing	K-plane	28.52	0.931	0.1009	2986
	HumanRF	27.84	0.846	0.0874	2852
	TiNeuVox	28.28	0.929	0.0956	5580
	ReRF	28.11	0.937	0.0688	879
	Ours(low)	27.84	0.931	0.0818	10.19
	Ours(high)	28.91	0.942	0.0669	64.92

Table 2. Per-scene results on ReRF dataset [9]. Values are averaged out over the number of frames in each scene.

		PSNR	SSIM	LPIPS	Size (KB)
flame samon	K-plane	30.57	0.925	0.2105	539
	NeRFPlayer	26.14	0.849	0.3790	2427
	Ours(low)	26.69	0.830	0.3486	26.70
	Ours(high)	28.05	0.872	0.2727	73.78
flame steak	K-plane	32.88	0.957	0.2021	539
	NeRFPlayer	27.36	0.867	0.3550	2427
	Ours(low)	30.11	0.891	0.3031	18.90
	Ours(high)	32.13	0.929	0.2295	56.12
coffee martini	K-plane	30.22	0.925	0.2113	539
	NeRFPlayer	32.05	0.938	0.2790	2427
	Ours(low)	26.28	0.822	0.3626	26.80
	Ours(high)	27.26	0.865	0.2890	76.37
cut roasted beef	K-plane	32.08	0.943	0.2196	539
	NeRFPlayer	31.83	0.928	0.2870	2427
	Ours(low)	29.60	0.887	0.3035	18.54
	Ours(high)	31.57	0.923	0.2374	56.13
cook spinach	K-plane	30.87	0.938	0.2212	539
	NeRFPlayer	32.06	0.930	0.2840	2427
	Ours(low)	29.40	0.882	0.3079	20.01
	Ours(high)	31.41	0.919	0.2398	60.00
sear steak	K-plane	31.69	0.955	0.2057	539
	NeRFPlayer	32.31	0.940	0.2720	2427
	Ours(low)	30.19	0.892	0.2994	17.81
	Ours(high)	32.18	0.931	0.2245	52.60

Table 3. Per-scene results on DyNeRF dataset [5]. Values are averaged out over the number of frames in each scene.

Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 1

[2] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 1

[3] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 1

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[5] Tianye Li, Miroslava Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, S. Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5511–5521, 2021. 2, 3

[6] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 1, 2

[7] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 1

[8] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19706–19716, 2023. 2

[9] Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, and Minye Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 76–87, 2023. 3

[10] L. Wang, Q. Hu, Q. He, Z. Wang, J. Yu, T. Tuytelaars, L. Xu, and M. Wu. Neural residual radiance fields for streamably free-viewpoint videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 76–87, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 2

[11] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020. 1, 2