Neural Radiance Fields for Dynamic View Synthesis using Local Temporal Priors

Rongsen Chen[1, 2], Junhong Zhao[1,2], Fang-Lue Zhang[1], Andrew Chalmers[1,2], Taehyun Rhee[1,2]

[1] School of Engineering and Computer Science, Victoria University of Wellington, Wellington, New Zealand
[2] Computational Media Innovation Centre (CMIC), Victoria University of Wellington, Wellington, New Zealand

CVM 2024 Wellington NZ

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

CMIC Computational Media Innovation Centre

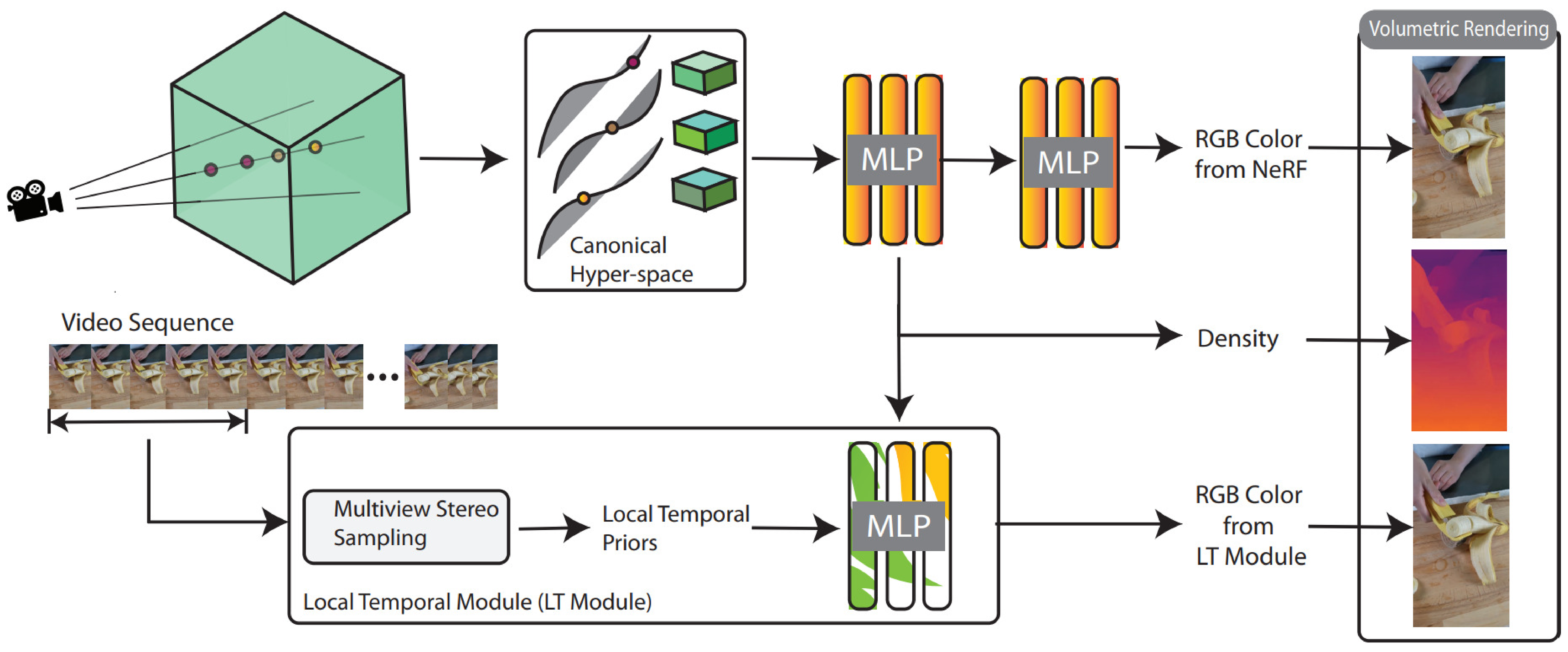
Introduction

- •Previous NeRF-based methods for non-rigid reconstruction although able to present visually appealing results, still often show visual artifacts such as blurry or incorrect geometry of an object.
- One of the causes is that previous work performs reconstruction directly on the entire video sequence. The global temporal information over the video sequence introduces noise to the network, often leading to a non-optimal canonical space representation of the dynamic scene.
- we present Local Temporal (LT) NeRF, a method to synthesize novel views of dynamic scenes using local temporal priors.

Contribution

- We present LT-NeRF, a novel view synthesis approach from monocular videos containing dynamic objects using NeRF enhanced by local temporal priors.
- We propose a novel LT module providing local temporal priors to improve the deformation field reconstruction and hyper-space encoding.
- We introduce two loss functions that take into account the temporal local information to supervise the deformation field and hyper-space encoding optimization.

Method



1. For each input image, our method samples 3D points in the observation volume and builds a hyper-space radiance field to associate 3D observation coordinates with hyper-space coordinates and predict their densities and colours.
2. We then leverage an MLP to estimate the colour of each spatial-temporal point by combining the colour information from temporally nearby frames and the learned geometric features.
3. Together with the original NeRF colour prediction, the two-colour outputs provide temporally global and local appearance information that are complementary to each other, leading to better geometry reconstruction from the radiance field.

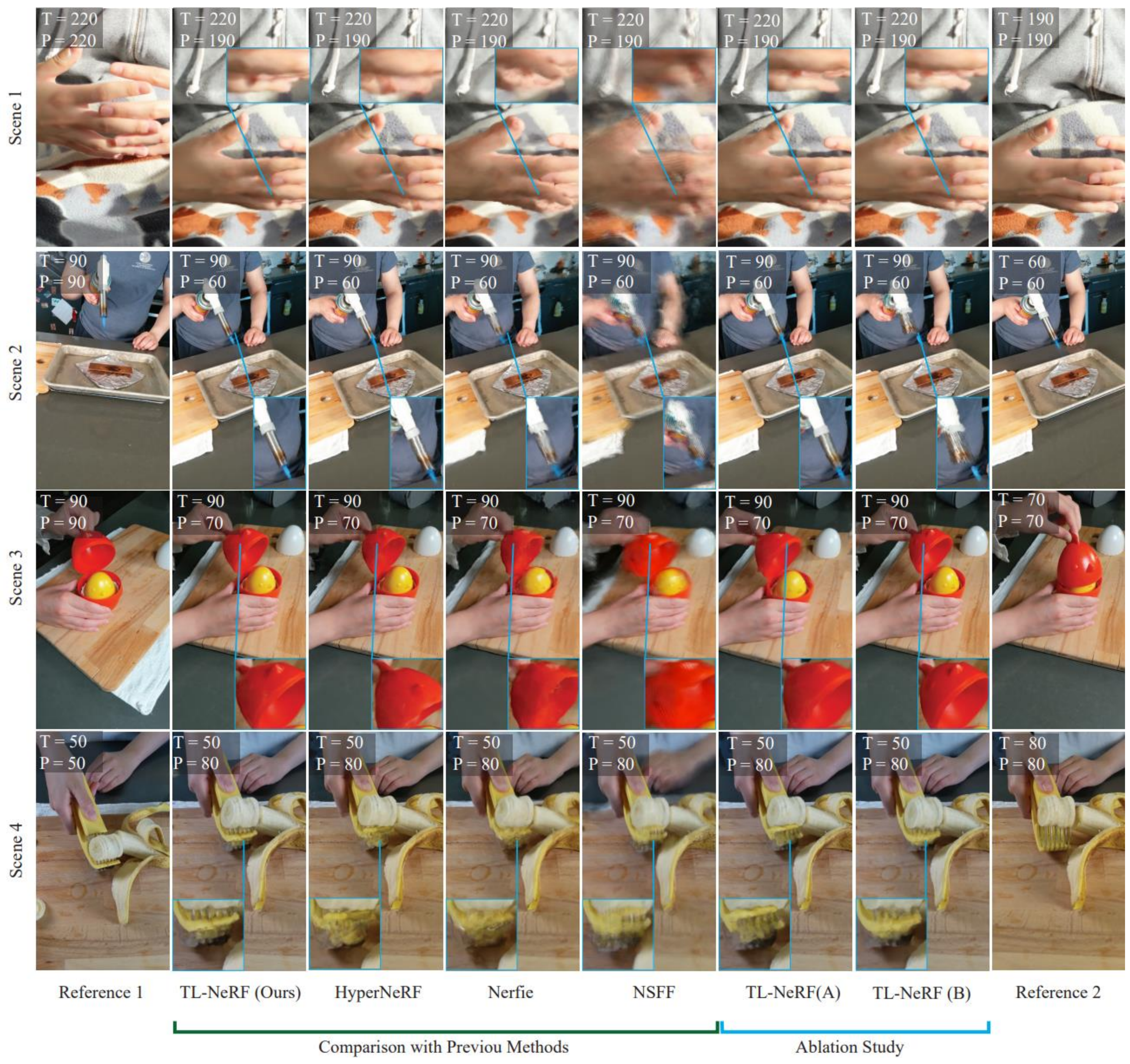
- Rongsen Chen: czn9678@gmail.com
- Junhong Zhao: j.zhao@vuw.ac.nz
- Fang-Lue Zhang: fanglue.zhang@vuw.ac.nz
- Andrew Chalmers: andrew.chalmers@vuw.ac.nz
- Taehyun Rhee: Taehyun.rhee@vuw.ac.nz

Quantitative Comparison

	Chicken (100 frames)			Banana (100 frames)			Chocolate (100 frames)			Hand (226 frames)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NSFF	24.41	0.8915	0.1329	27.78	0.9195	0.0992	27.58	0.9651	0.0624	29.87	0.9766	0.074
Nerfie	27.55	0.9485	0.1037	30.15	0.9609	0.0821	26.27	0.9528	0.0618	27.55	0.9594	0.0713
HyperNeRF	26.89	0.9390	0.0867	29.07	0.9594	0.0668	28.09	0.9768	0.0413	27.07	0.9586	0.0658
LT-NeRF (Ours)	29.67	0.9701	0.0554	32.56	0.9802	0.0382	29.05	0.9815	0.0240	32.35	0.9888	0.0568
LT-NeRF (A)	25.55	0.9272	0.0983	31.35	0.9697	0.0492	28.09	0.9758	0.0351	30.65	0.9697	0.0615
LT-NeRF (B)	29.65	0.9695	0.0609	31.98	0.9767	0.0369	29.09	0.9815	0.0283	32.16	0.9888	0.0593

	VRIG-Chicken (164 frames)			VRIG-3D Printer (207 frames)			VRIG-Broom (197 frames)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	19.9	0.777	0.325	20.7	0.780	0.357	19.9	0.653	0.692
NV	17.6	0.615	0.336	16.2	0.665	0.330	17.7	0.623	0.360
NSFF	26.9	0.944	0.106	27.7	0.947	0.125	26.1	0.871	0.284
Nerfie	26.7	0.943	0.078	20.6	0.830	0.108	19.2	0.567	0.325
HyperNeRF	26.9	0.948	0.079	20.0	0.821	0.111	19.3	0.591	0.269
LT-NeRF (Ours)	26.0	0.904	0.102	23.1	0.885	0.097	21.9	0.742	0.217

Qualitative Comparison



Conclusions

Our novel LT module provides the local temporal priors using multi-view stereo sampling and has shown improved performance on the tested dataset.

Limitation

Our LT-NeRF approach may face challenges in certain extreme cases, such as videos with fast-moving objects, scenes with challenging camera angles, and scenarios were differentiating between the dynamic object and the background is difficult.

Acknowledgement

This work was supported by the Entrepreneurial University Programme from the Tertiary Education Commission of New Zealand.