ZeST-NeRF: Using temporal aggregation for Zero-Shot Temporal NeRFs - Supplementary

Violeta Menéndez González[1,2]
v.menendezgonzalez@surrey.ac.uk
Andrew Gilbert[1]
a.gilbert@surrey.ac.uk
Graeme Phillipson[2]
graeme.phillipson@bbc.co.uk
Stephen Jolly[2]
stephen.jolly@bbc.co.uk
Simon Hadfield[1]
s.hadfield@surrey.ac.uk

[1] Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey
Guildford, UK

[2] BBC R&D
MediaCityUK
Salford, UK

1 Implementation details

We use COLMAP [10] to generate camera intrinsics and extrinsics at each frame while masking features from regions associated with dynamic objects [8] using off-the-shelf instance segmentation [4]. We extract deep image features from the selected frames using a 2D CNN network with 32 channels (first section of Table 1). These features are used to construct the plane sweep volume [2] using 128 depth planes. These sweep volumes are then aggregated into a variance-based cost volume. This is then processed into the *geometry* and *motion* volumes as defined by the 3D CNN architecture on the second section of Table 1. These volumes have the same architecture, only differing in the number of input channels ($K = 8$ key-frames and $N = 4$ neighbours, respectively). The *geometry* and *motion* volumes do not share their weights.

For the NeRF MLPs, we follow a similar setup to the original case [9]. We sample 128 points along each ray, with a ray batch of 1024. We also have two separate networks for the static and dynamic parts, which do not share weights. We append the normalised time indices in NSFF [8] to our dynamic network inputs. The MLP networks return the estimated colour c and density σ, as well as blending weights b in the case of the Static MLP, and 3D scene flow f and occlusion weights w in the case of the Dynamic MLP. We use an Adam optimiser [6] with a learning rate of $5e-4$. We use positional encoding (PE) [9] for the 3D location and viewing direction before feeding them into the networks. For more detailed information about the architecture, refer to the Table 2.

Table 1: **Encoding volumes architecture**: g/m denote the geometry and motion 3D features respectively. **k** is the kernel size, **s** is the stride, **d** is the kernel dilation, and **chns** shows the number of input and output channels for each layer. We denote CBR2D/CBR3D/CTB3D to be ConvBnReLU2D, ConvBnReLU3D, and ConvTransposeBn3D layer structure respectively.

	Layer	**k**	**s**	**d**	**chns**	input
2D CNN	CBR2D_0	3	1	1	$3/8$	I
	CBR2D_1	3	1	1	$8/8$	CBR2D_0
	CBR2D_2	5	2	2	$8/16$	CBR2D_1
	CBR2D_3	3	1	1	$16/16$	CBR2D_2
	CBR2D_4	3	1	1	$16/16$	CBR2D_3
	CBR2D_5	5	2	2	$16/32$	CBR2D_4
	CBR2D_6	3	1	1	$32/32$	CBR2D_5
	$E = \text{CBR2D}_7$	3	1	1	$32/32$	CBR2D_6
3D CNN	CBR3D_0	3	1	1	$32+(K/N)*3/8$	E,I
	CBR3D_1	3	2	1	$8/16$	CBR3D_0
	CBR3D_2	3	1	1	$16/16$	CBR3D_1
	CBR3D_3	3	2	1	$16/32$	CBR3D_2
	CBR3D_4	3	1	1	$32/32$	CBR3D_3
	CBR3D_5	3	2	1	$32/64$	CBR3D_4
	CBR3D_6	3	1	1	$64/64$	CBR3D_5
	CTB3D_0	3	2	1	$64/32$	CBR3D_6
	CTB3D_1	3	2	1	$64/32$	$\text{CTB3D}_0 + \text{CBR3D}_4$
	CTB3D_2	3	2	1	$64/32$	$\text{CTB3D}_1 + \text{CBR3D}_2$
	$g/m = \text{CTB3D}_3$	3	2	1	$64/32$	$\text{CTB3D}_2 + \text{CBR3D}_0$

Table 2: **MLPs architecture**: g/m denote the geometry and motion 3D features respectively. k and n are the original colours of the K key-frames and N neighbouring frames, that are concatenated to the inputs. **chns** shows the number of input and output channels for each layer. We denote LR to be LinearReLU layer structure. PE refers to the positional encoding as used in [9].

	Layer	**chns**	input
Static MLP	PE_0	3/63	x
	LR_0	8+K*3/256	g, k
	LR_1	63/256	PE
	LR_{i+1}	256/256	LR_i+LR_0
	σ	256/1	LR_6
	b	256/1	LR_6
	PE_1	3/27	d
	LR_7	27+256/256	PE_1,LR_6
	c	256/3	LR_7
Temporal MLP	PE_0	4/63	x, t
	LR_0	8+N*3/256	m, n
	LR_1	63/256	PE
	LR_{i+1}	256/256	LR_i+LR_0
	σ	256/1	LR_6
	f	256/6	LR_6
	w	256/2	LR_6
	PE_1	3/27	d
	LR_7	27+256/256	PE_1,LR_6
	c	256/3	LR_7

2 Evaluation of accuracy

In order to assess the performance of our model, we employ a range of widely recognized metrics that evaluate various aspects of an image. To measure image quality we make use of the Peak Signal-To-Noise Ratio (PSNR) [5] and the Structural SIMilarity (SSIM) [11] index. PSNR serves as an indicator of the overall consistency of pixels, while SSIM gauges the coherency of local structures. We define PSNR as

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_C^2}{MSE\left(\hat{C}^b(\mathbf{r}), C(\mathbf{r})\right)}\right) \tag{1}$$

$$MSE\left(\hat{C}^b(\mathbf{r}), C(\mathbf{r})\right) = \frac{1}{N}\sum_{\mathbf{r}}[\hat{C}^b(\mathbf{r}) - C(\mathbf{r})]^2 \tag{2}$$

where MAX_C is the maximum possible input value, and $MSE\left(\hat{C}^b(\mathbf{r}), C(\mathbf{r})\right)$ represents the per-pixel Maximum Squared Error between the predicted colour $\hat{C}^b(\mathbf{r})$ at ray $\mathbf{r}$, and the original colour $C(\mathbf{r})$, in a batch of N rays.

On the other hand, SSIM is given by

$$SSIM(\hat{C}^b, C) = \frac{(2\mu_{\hat{C}^b}\mu_C + k_1)(2\sigma_{\hat{C}^b}\sigma_C + k_2)}{(\mu_{\hat{C}^b}^2 + \mu_C^2 + k_1)(\sigma_{\hat{C}^b}^2 + \sigma_C^2 + k_2)} \tag{3}$$

where $k_1 = 0.01^2$ and $k_2 = 0.03^2$ are variables to stabilise the operation. We use a window size of 5 for the Gaussian kernel to smooth the images.

It is worth noting that these metrics assume independence among pixels, which can result in favourable scores for visually inaccurate outcomes. Consequently, we also incorporate the application of a Learned Perceptual Image Patch Similarity (LPIPS) [13] metric, which endeavours to capture human perception by leveraging deep features. We use the default settings for the implementation based on AlexNet [7].

For qualitative results, see Figure 1 in Section 3.

3 Further results

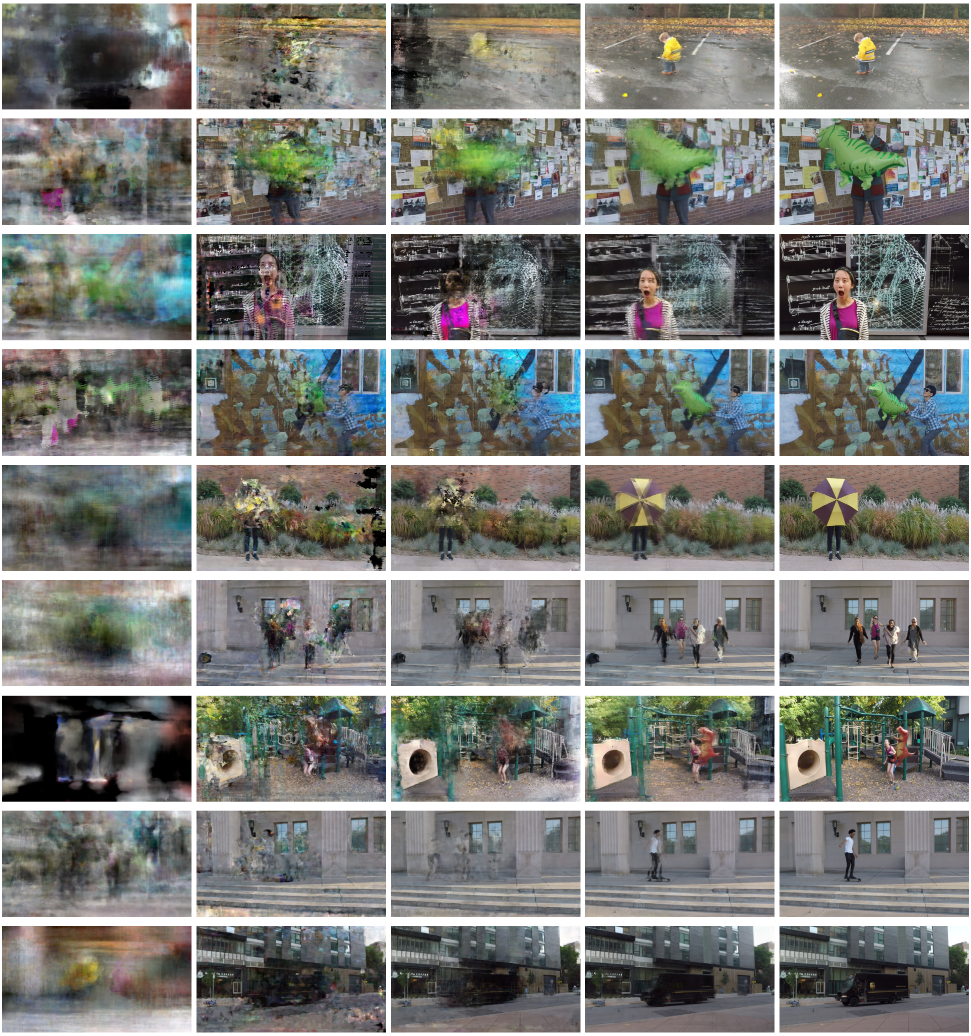

Figure 1: **Qualitative results** on the Dynamic Scenes dataset [12]

References

[1] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. DeepStereo: Learning

to Predict New Views from the World's Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Violeta Menéndez González, Andrew Gilbert, Graeme Phillipson, Stephen Jolly, and Simon Hadfield. SVS: Adversarial refinement for sparse novel view synthesis. In *Proceedings of the 33rd British Machine Vision Conference (BMVC)*, 2022.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[5] Q. Huynh-Thu and M. Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 2008.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2014.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, January 2012.

[8] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[9] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[10] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] Wang, Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, April 2004.

[12] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel View Synthesis of Dynamic Scenes With Globally Coherent Depths From a Monocular Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.