Visual Learning And Reasoning

The Hong Kong Polytechnic University
香港理工大學

International Conference On Machine Learning

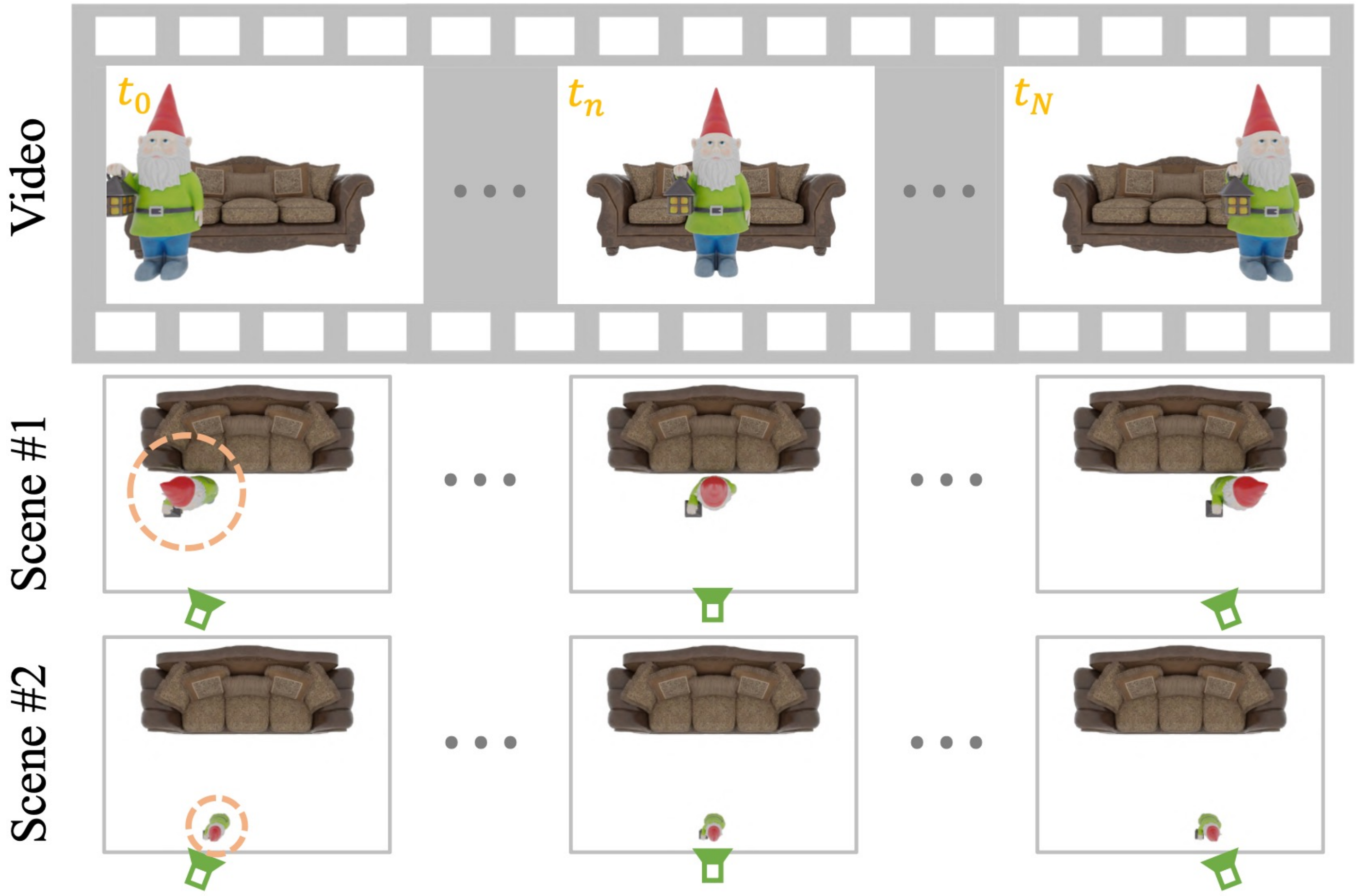
OSN: Infinite Representations of Dynamic 3D Scenes from Monocular Videos

Ziyang Song, Jinxi Li, Bo Yang

vLAR Group, The Hong Kong Polytechnic University

Introduction

Our task: Reconstruct dynamic 3D scenes from monocular videos



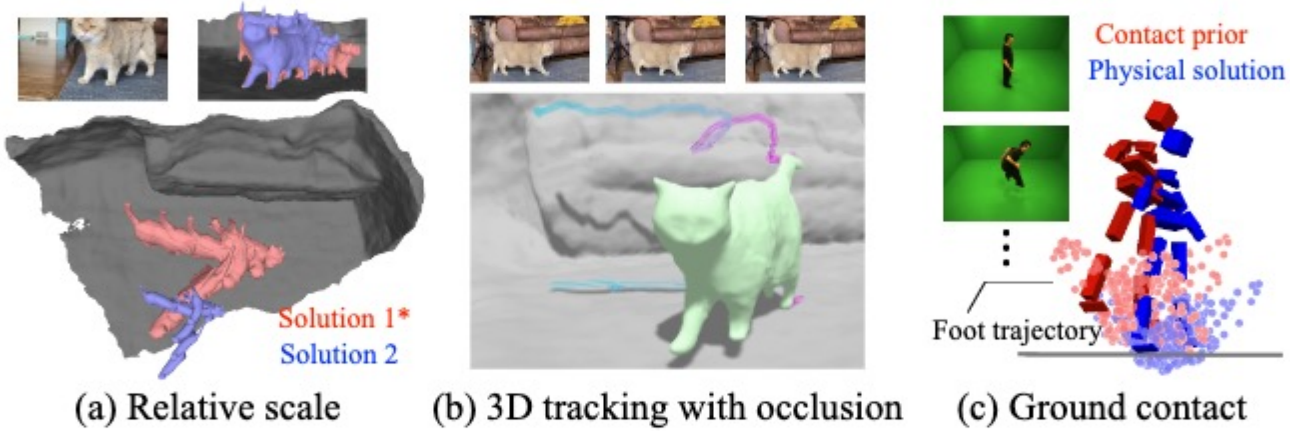
Highly ill-posed problem:
Many correct 3D scenes correspond to the video

Introduction

Prior works

Monocular depths [1][2]

Physical constraints [3]



Finding a single solution:

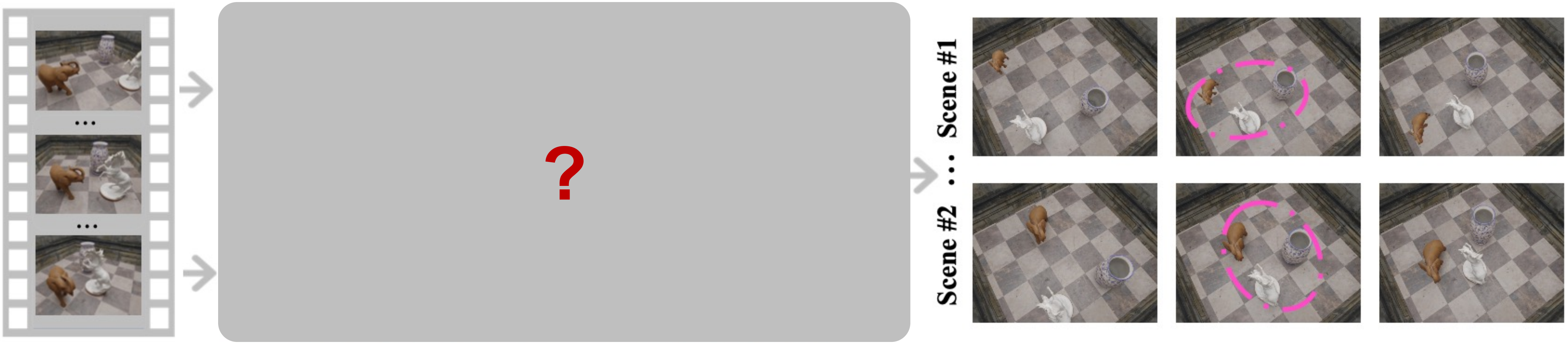
- Not general enough
- Additional constraints may not always be reliable

[1] Z. Li, S. Niklaus, N. Snavely, et al. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. CVPR, 2021.
[2] C. Gao, A. Saraf, J. Kopf, et al. Dynamic View Synthesis from Dynamic Monocular Video. ICCV, 2021.
[3] G. Yang, S. Yang, J. Z. Zhang, et al. PPR: Physically Plausible Reconstruction from Monocular Videos. ICCV, 2023.

Introduction

Our goal:

Learn **all** plausible 3D scene configurations that **match the input video**

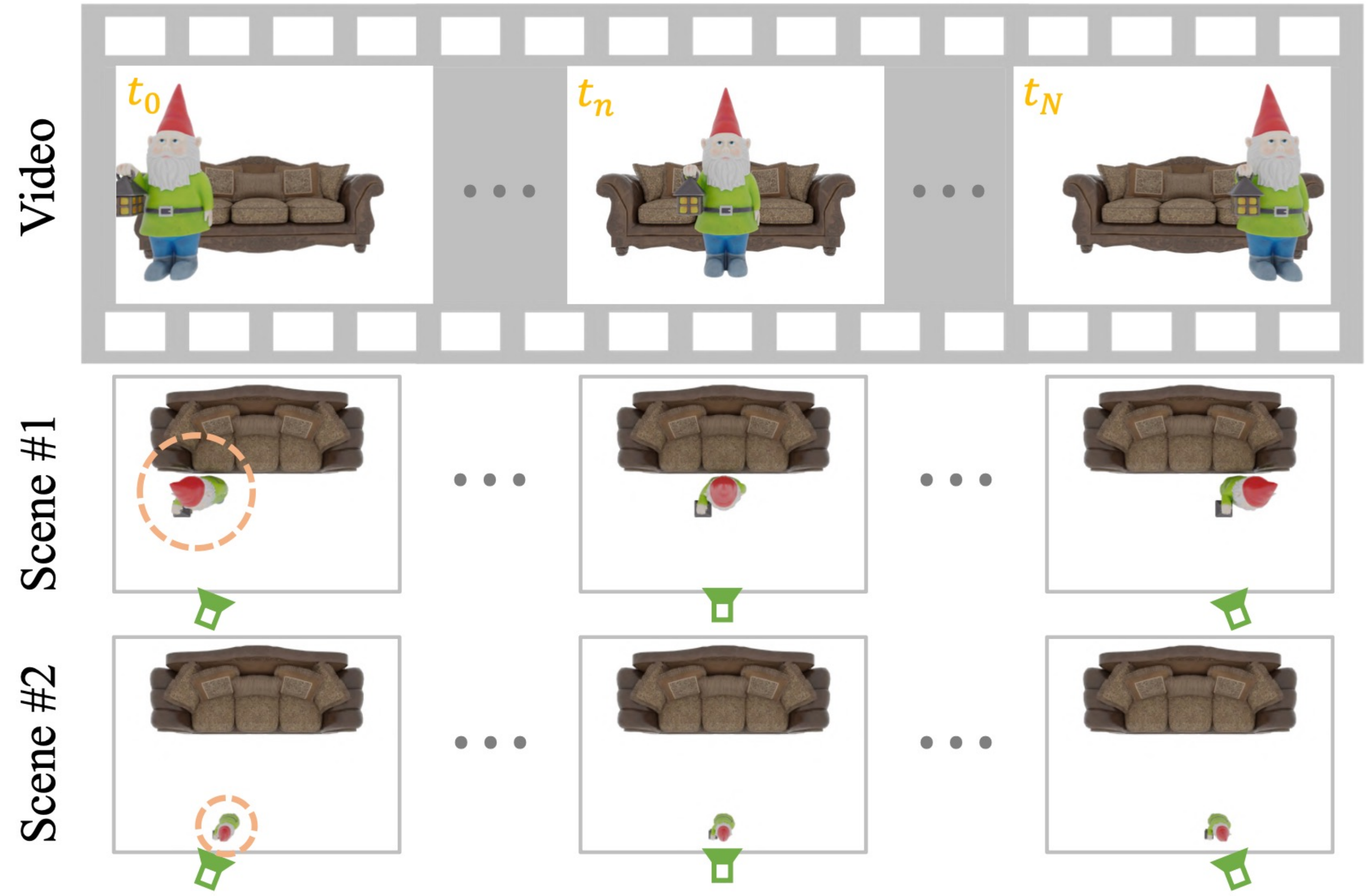


Q1: How to represent?

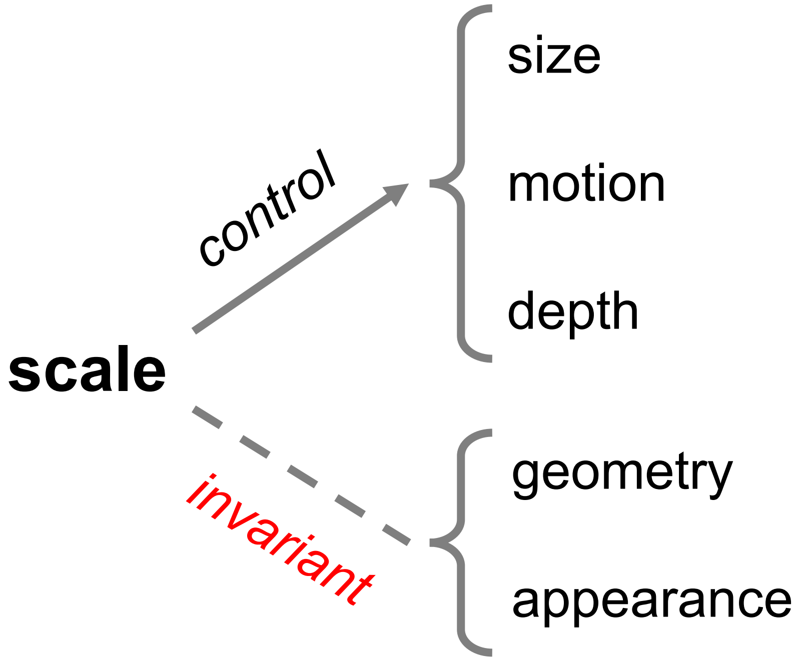
Q2: How to learn?

OSN

How to represent?

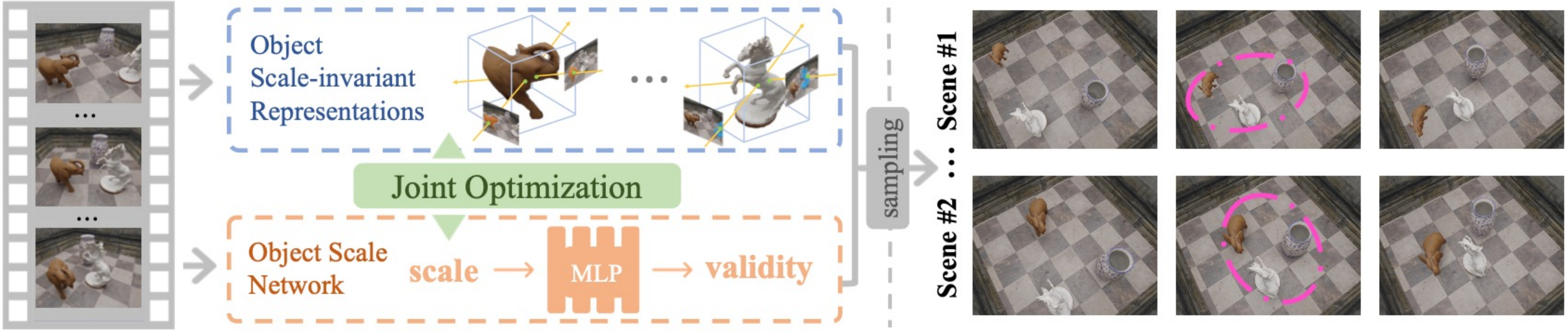


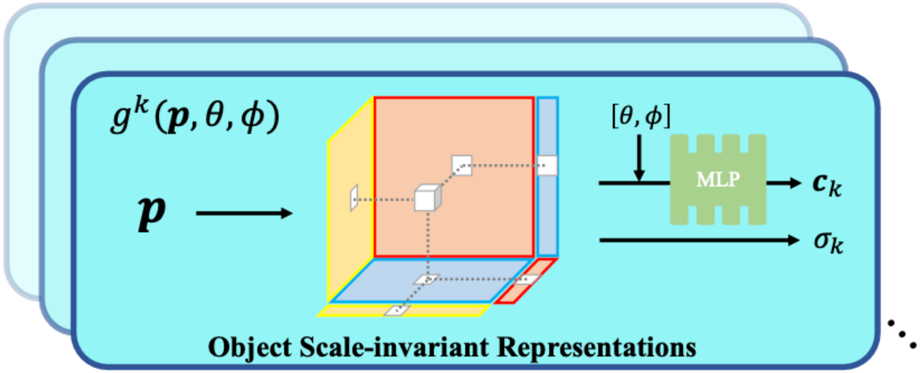
For a rigid object:

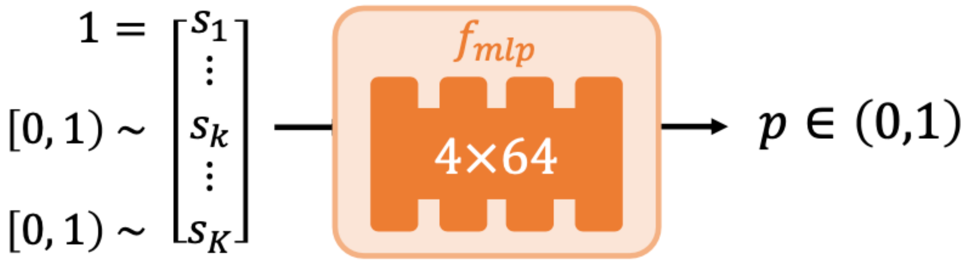


OSN

Framework







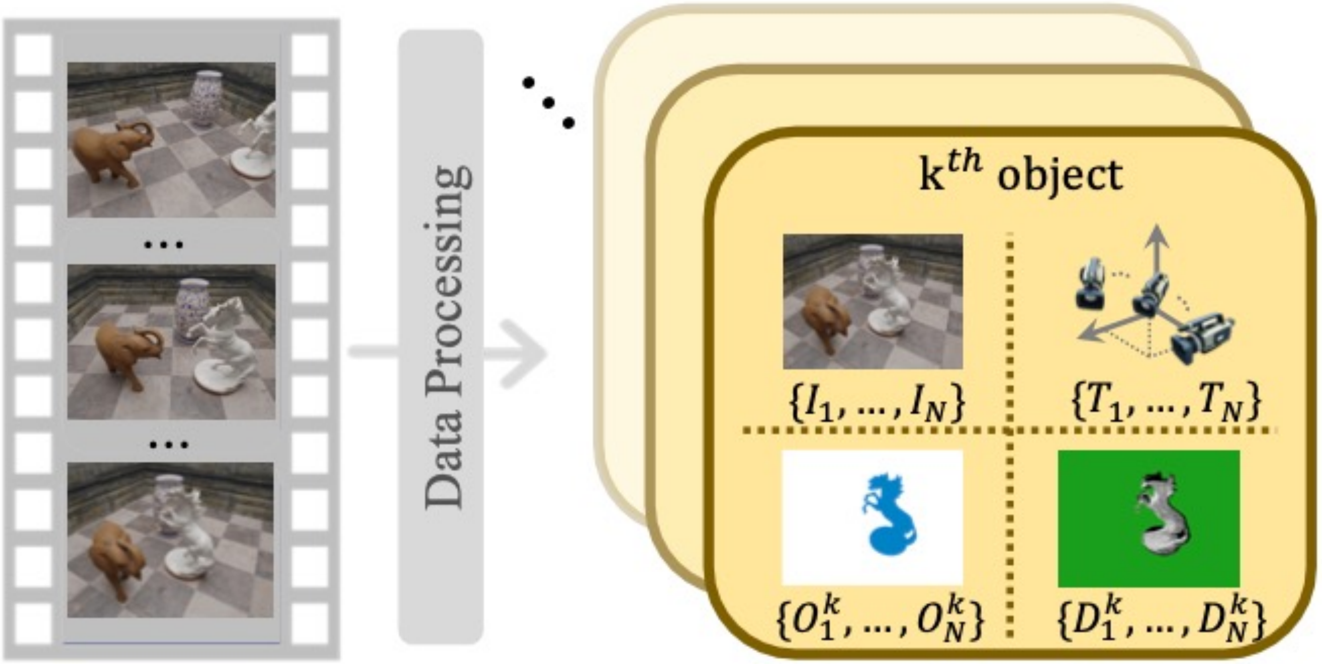
Object Scale-invariant Representations: TensoRF[1]

Object Scale Network: MLP

[1] A. Chen, Z. Xu, A. Geiger, et al. TensoRF: Tensorial Radiance Fields. ECCV, 2022.

OSN

How to learn?



Preprocessing:

- SAM [1] & TAM [2]
- RAFT [3]
- SfM [4]

Available information:

- RGB
- segmentation masks
- camera-to-object poses
- per-object relative depths

[1] A. Kirillov, E. Mintun, N. Ravi, et al. Segment Anything. ICCV, 2023.
[2] J. Yang, M. Gao, Z. Li, et al. Track Anything: Segment Anything Meets Videos. arXiv:2304.11968, 2023.
[3] Z. Teed, and J. Deng. RAFT: Recurrent All Pairs Field Transforms for Optical Flow. ECCV, 2020.
[4] J. L. Schonberger, and J.-M. Frahm. Structure-from-Motion Revisited. CVPR, 2016.

OSN

How to optimize object representations?

Sample **one valid** scale combination

Composite rendering [1] → *Scaled* composite rendering →

$$\ell_{rgb}^{scene} = \sum_{\mathbf{r}^k} ||\mathbf{c}(\mathbf{r}^k) - \bar{\mathbf{c}}(\mathbf{r}^k)||$$

$$\ell_{depth}^{scene} = \sum_{\mathbf{r}^k} ||d(\mathbf{r}^k) - s_k * \bar{d}(\mathbf{r}^k)||$$

$$\ell_{seg}^{scene} = \sum_{\mathbf{r}^k} CE\Big(\mathbf{o}(\mathbf{r}^k), \bar{\mathbf{o}}(\mathbf{r}^k)\Big)$$

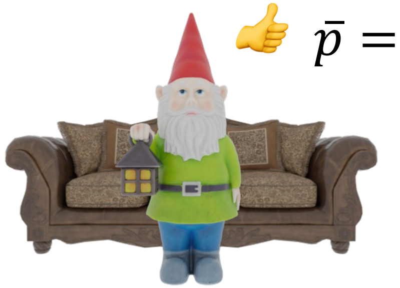
How to optimize object scale network?

Sample **many** (valid / invalid) scale combinations

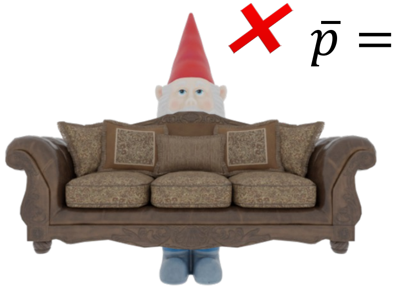
pseudo GT = **segmentation** (inter-object occlusion) correctness

$$\bar{p}^h = \sum \Big(|\mathbf{o}^h(\mathbf{r}^k)| * \bar{\mathbf{o}}(\mathbf{r}^k)\Big) \rightarrow 0/1$$

$$\ell_{bce} = \sum_{\mathbf{r}^k} \Big(\sum_h BCE(p^h, \bar{p}^h)\Big)$$



$\bar{p} = 1$



$\bar{p} = 0$

[1] C. Song, G. Yang, K. Deng, et al. Total-Recon: Deformable Scene Reconstruction for Embodied View Synthesis. ICCV, 2023.

OSN

Soft Z-buffer rendering

Rendering under H scale combinations is time-consuming

$[s_1, \cdots, s_K]$

Render (up-to-scale) per-object **depth** independently

$s_1 \cdot d^1(\boldsymbol{r})$

$s_2 \cdot d^2(\boldsymbol{r})$

Segmentation determined by scaled depth

$$\mathbf{o}^h(\mathbf{r}^k) = \left[\cdots, \frac{e^{-\left(s_k^h * d^k(\mathbf{r}^k)\right)}}{\sum_{k=1}^{K} e^{-\left(s_k^h * d^k(\mathbf{r}^k)\right)}}, \cdots\right]$$

Comparison

- ***Scaled composite rendering:*** $H \times$ 3D volume rendering
- ***Soft Z-buffer rendering:*** $1 \times$ 3D volume rendering + $H \times$ 2D image blending

OSN

Joint training procedure

Stage 1 – Bootstrapping per-object representations

Stage 2 – Alternative optimization

Optimize object representations ⇄ Optimize object scale network

Results

Input Monocular Video

Multiple Possible Dynamic 3D Scenes

S^1

S^3

S^2

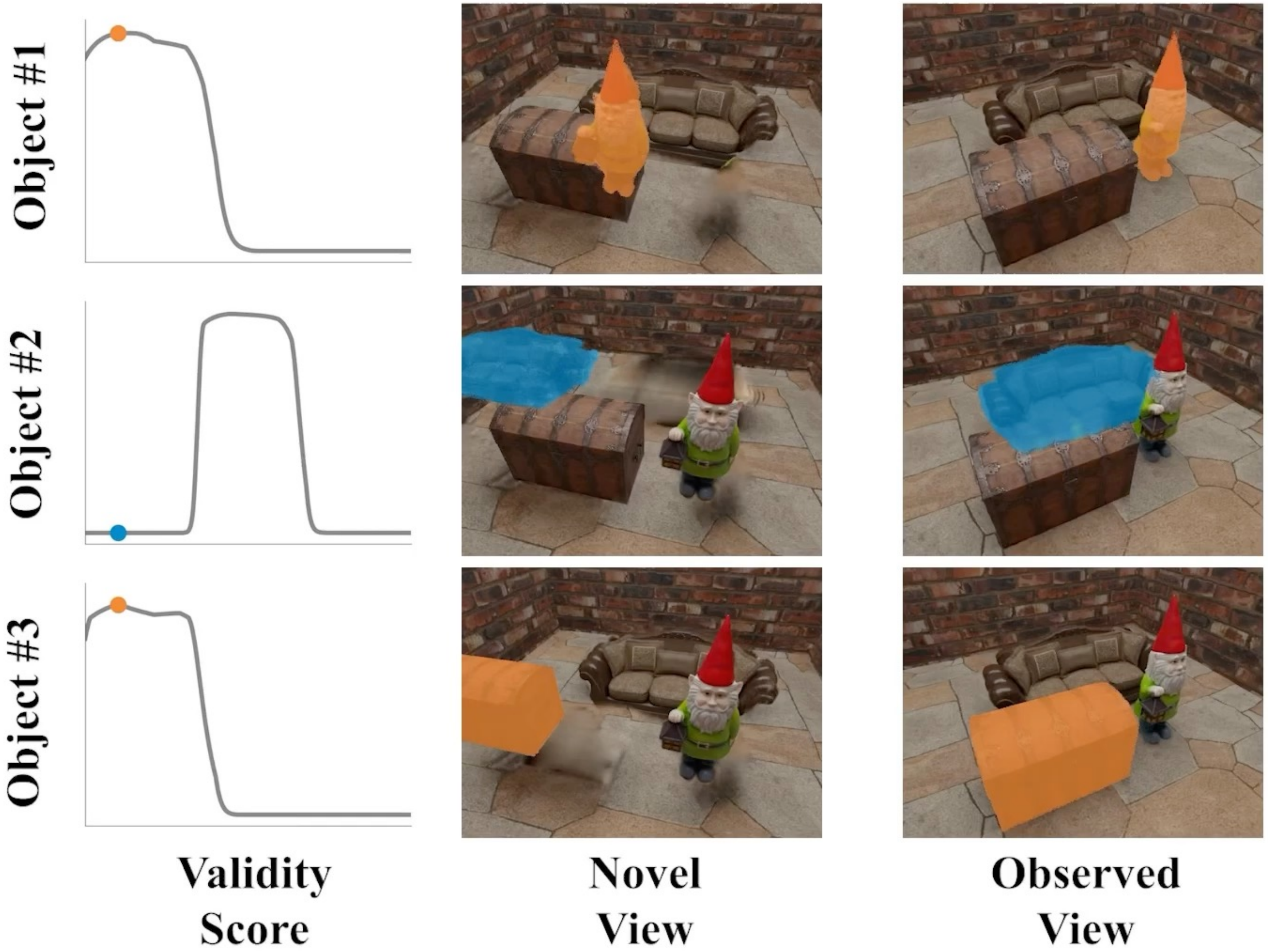
S^4

Observed View Novel View

Observed View Novel View

Results

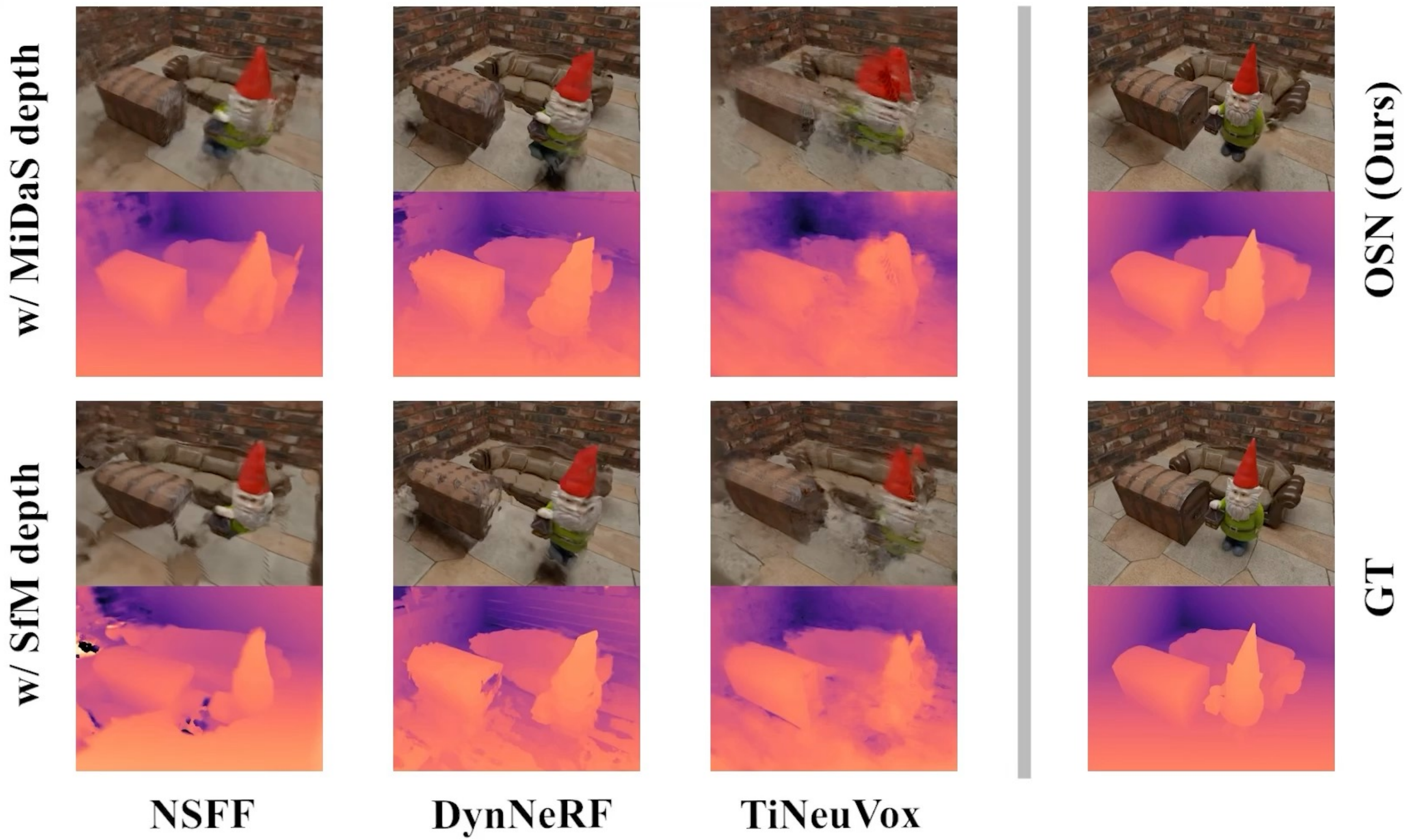
Analysis of validity scores



Fixed timestamp = 0
Change object scale

Results

Dynamic novel view synthesis



Ours: the best of 1000 samples

Results

Dynamic novel view synthesis

Table 1. Quantitative results of all methods for dynamic novel view synthesis on three datasets. The methods are trained with different depth supervision: 1) w/o depth, 2) w/ MiDaS depth, and 3) w/ per-object SfM depth.

		Dynamic Indoor Scene Dataset				Oxford Multimotion Dataset			NVIDIA Dynamic Scene Dataset		
Depth Sup.	Method	PSNR↑	SSIM↑	LPIPS↓	SSIMAE↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1)	NSFF(Li et al., 2021)	21.428	0.720	0.313	0.378	16.687	0.616	0.249	21.766	0.669	0.229
	DynNeRF(Gao et al., 2021)	21.479	0.752	0.277	0.417	16.858	0.627	0.244	25.705	0.827	0.117
	TiNeuVox(Fang et al., 2022)	21.705	0.655	0.306	0.484	16.433	0.613	0.325	22.922	0.618	0.262
	HexPlane(Cao & Johnson, 2023)	18.637	0.581	0.480	0.962	17.084	0.631	0.221	20.169	0.555	0.286
2)	NSFF(Li et al., 2021)	20.900	0.698	0.349	0.494	17.094	0.623	0.244	27.459	0.861	0.075
	DynNeRF(Gao et al., 2021)	22.272	0.767	0.257	0.309	16.521	0.622	0.259	29.452	**0.895**	0.054
	TiNeuVox(Fang et al., 2022)	23.288	0.698	0.269	0.329	18.508	0.668	0.197	23.029	0.621	0.193
	HexPlane(Cao & Johnson, 2023)	17.968	0.528	0.535	1.395	15.843	0.576	0.338	19.312	0.471	0.334
3)	NSFF(Li et al., 2021)	21.280	0.684	0.347	0.467	17.093	0.616	0.245	23.733	0.733	0.194
	DynNeRF(Gao et al., 2021)	21.421	0.742	0.296	0.509	16.786	0.624	0.281	24.498	0.771	0.176
	TiNeuVox(Fang et al., 2022)	22.197	0.685	0.285	0.368	18.043	0.670	0.208	22.691	0.591	0.215
	HexPlane(Cao & Johnson, 2023)	20.217	0.623	0.373	0.458	17.137	0.631	0.203	23.220	0.720	0.150
	OSN(Ours)	**25.984**	**0.861**	**0.115**	**0.094**	**19.671**	**0.695**	**0.155**	**29.588**	0.892	**0.053**
2)+3)	Total-Recon(Song et al., 2023)	24.695	0.841	0.128	0.137	18.331	0.655	0.173	27.822	0.880	0.059

Results

Dynamic novel view synthesis -- multiple GT

Table 2. Quantitative results of all methods for dynamic novel view synthesis on synthetic "Gnome House" scene with 50 different ground truth scale combinations. The average performance along with standard deviations on 50 groups of ground truths are reported. The methods are trained with different depth supervision: 1) w/o depth, 2) w/ MiDaS depth, and 3) w/ per-object SfM depth.

		50 Ground Truth Scenes of Gnome House			
Depth Sup.	Method	PSNR↑	SSIM↑	LPIPS↓	SSIMAE↓
1)	NSFF(Li et al., 2021)	19.088±1.514	0.636±0.026	0.385±0.029	0.559±0.183
	DynNeRF(Gao et al., 2021)	18.846±1.227	0.645±0.023	0.380±0.027	0.540±0.156
	TiNeuVox(Fang et al., 2022)	18.361±1.159	0.539±0.026	0.414±0.033	0.600±0.140
	HexPlane(Cao & Johnson, 2023)	16.762±0.130	0.420±0.002	0.708±0.005	1.688±0.098
2)	NSFF(Li et al., 2021)	18.993±1.485	0.592±0.024	0.465±0.027	0.582±0.180
	DynNeRF(Gao et al., 2021)	18.759±1.398	0.639±0.029	0.378±0.032	0.579±0.194
	TiNeuVox(Fang et al., 2022)	18.978±1.249	0.560±0.028	0.394±0.035	0.619±0.159
	HexPlane(Cao & Johnson, 2023)	17.325±0.605	0.434±0.015	0.626±0.019	1.993±0.119
3)	NSFF(Li et al., 2021)	18.214±0.948	0.492±0.016	0.536±0.020	0.776±0.137
	DynNeRF(Gao et al., 2021)	18.767±1.270	0.639±0.026	0.382±0.029	0.554±0.160
	TiNeuVox(Fang et al., 2022)	18.776±1.155	0.556±0.027	0.396±0.033	0.553±0.154
	HexPlane(Cao & Johnson, 2023)	18.464±0.767	0.492±0.019	0.480±0.025	0.660±0.130
	OSN(Ours)	**22.940±1.004**	**0.784±0.022**	**0.160±0.021**	**0.125±0.078**
2)+3)	Total-Recon(Song et al., 2023)	18.768±1.535	0.666±0.032	0.295±0.046	0.612±0.212

Conclusion & Future Directions

Our contributions:

- First work to represent 3D scenes in many ways from a monocular video
- An object scale network with a joint optimization method
- Effectiveness on synthetic and real-world datasets

Future directions:

- Infinite solutions for monocular dynamic scenes with deformable objects

Thanks

paper & code: Coming soon!

contact: ziyang.song@connect.polyu.hk