GaussianPrediction: Dynamic 3D Gaussian Prediction for Motion Extrapolation and Free View Synthesis
Supplementary Material

Boming Zhao*
bmzhao@zju.edu.cn
Zhejiang University
Hangzhou, China

Yuan Li*
yuan_li@zju.edu.cn
Zhejiang University
Hangzhou, China

Ziyu Sun
sunzy2121@mails.jlu.edu.cn
Jilin University
Changchun, China

Lin Zeng
22251265@zju.edu.cn
Zhejiang University
Hangzhou, China

Yujun Shen
shenyujun0302@gmail.com
Ant Group
Hangzhou, China

Rui Ma
ruim@jlu.edu.cn
Jilin University
Changchun, China

Yinda Zhang
yindaz@google.com
Google Inc.
Mountain View, USA

Hujun Bao
bao@cad.zju.edu.cn
Zhejiang University
Hangzhou, China

Zhaopeng Cui†
zhpcui@gmail.com
Zhejiang University
Hangzhou, China

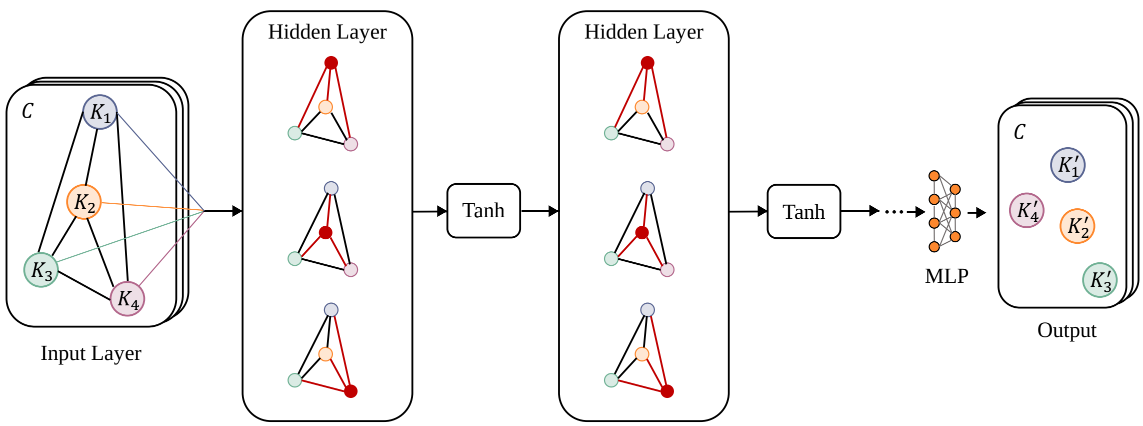

Figure A: GCN network architecture.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; **Rendering**.

KEYWORDS

novel view synthesis, dynamics modeling, future prediction

*Boming Zhao and Yuan Li contributed equally to this work. The authors from Zhejiang University are also affiliated with the State Key Laboratory of CAD&CG.
†Corresponding author.

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657417

In this supplementary material, we describe more details of our method in Sec. A. We also conduct more experiments in Sec. B.

A MORE DETAILS ON GCN-BASED MOTION PREDICTION

We present our GCN network architecture as shown in Fig. A. We use two separate GCN networks to predict the 3D positions of key points and the rotation represented by quaternions, respectively. We model the N key points of the entire scene as a fully connected graph, where the connection strength between each key point is learned during training, which is captured by the weighted adjacency matrix $A \in \mathbb{R}^{N \times N}$. Then a graph convolutional layer g takes a set of weights $W^g \in \mathbb{R}^{F \times \hat{F}}$ and the features $X^g \in \mathbb{R}^{N \times F}$ from the previous layer and outputs the feature for the next layer as follows:

$$X^{g+1} = \sigma(AX^g W^g), \tag{1}$$

where σ represents the activation function.

Table A: Quantitative results comparison for motion extrapolation with 4D-Gs [Wu et al. 2023] and Deform-GS [Yang et al. 2023] on Hyper-NeRF real-dataset. Best results are highlighted as first, second.

Method	CHICKEN (23 images)		CUT LEMON (83 images)		SPLIT COOKIE (27 images)		3D PRINTER (42 images)		AVERAGE	
	PSNR(↑)	MS-SSIM(↑)	PSNR(↑)	MS-SSIM(↑)	PSNR(↑)	MS-SSIM(↑)	PSNR(↑)	MS-SSIM(↑)	PSNR(↑)	MS-SSIM(↑)
4D-GS	17.7	.659	20.2	.688	18.1	.623	16.0	.460	18.5	.619
Deform-GS	17.8	.686	19.2	.591	17.3	.590	15.7	.445	17.9	.568
Ours	18.0	.675	20.0	.688	18.4	.673	16.2	.484	18.6	.635

During training, we use the key points information from each frame in the training data as pseudo ground truth. Our model takes 10 frames of 3D positions and rotations of key points as inputs and predicts the subsequent moment information of key points. In inference, we employ a sliding window strategy to predict the key points of information across multiple frames. Specifically, the GCN takes the last 10 frames of the training data as input to predict one frame. Subsequently, the predicted frame, along with the previous 9 frames, is used as the new input to predict the next frame, and so on. This method enables us to render long predictive sequences.

However, during our experiments, we found that using a K-layer GCN to directly predict the 3D positions of key points results in the loss of knowledge acquired during training after predicting several frames, tending towards meaningless linear motion. Inspired by SIMLPE [Guo et al. 2023], we employed a two-layer tiny MLP to decode the features from GCN, effectively capturing the motion characteristics. Additionally, due to the limited training data available for our model, we have introduced progressively decreasing noise to the input 3D positions and rotations with each training epoch, to prevent the network from overfitting to the training set.

B MORE EXPERIMENTS

We show prediction quantitative evaluations of the real-world HyperNeRF dataset compared with 4D-GS [Wu et al. 2023] and Deforma-GS [Yang et al. 2023] in Table A. Note that in real-world scene-level datasets, the predicted results cannot be perfectly aligned with the ground truth images due to ill camera poses and inaccurate timestamps, which makes the quantitative comparison less meaningful than the qualitative comparison. Nevertheless, we still achieved SOTA results. Our video includes all motion extrapolation results, demonstrating that our method better captures and extrapolates motion patterns in all scenarios. Please refer to our video for more details.

REFERENCES

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2023. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4809–4819.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528* (2023).

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2023. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101* (2023).