GenFusion: Closing the loop between Reconstruction and Generation via Videos

Sibo Wu, Congrong Xu, Bonbon Huang, Andreas Geiger, Anpei Chen

12/16/2024

Self-introduction

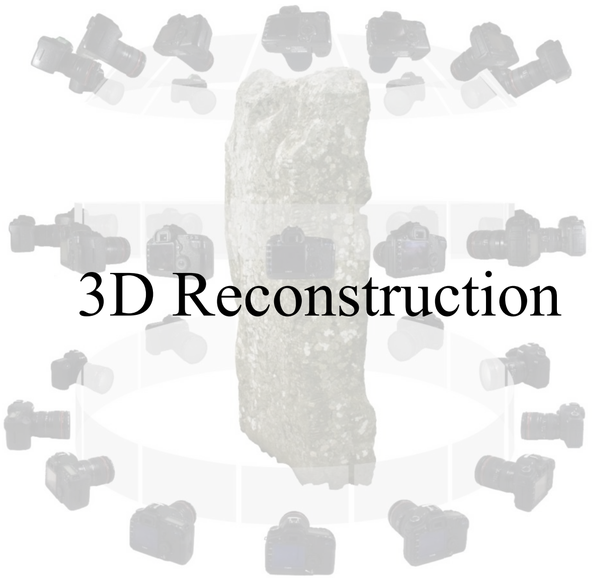
Sibo Wu

Master student @ Technical University of Munich

Research interest: 3D reconstruction and generative model

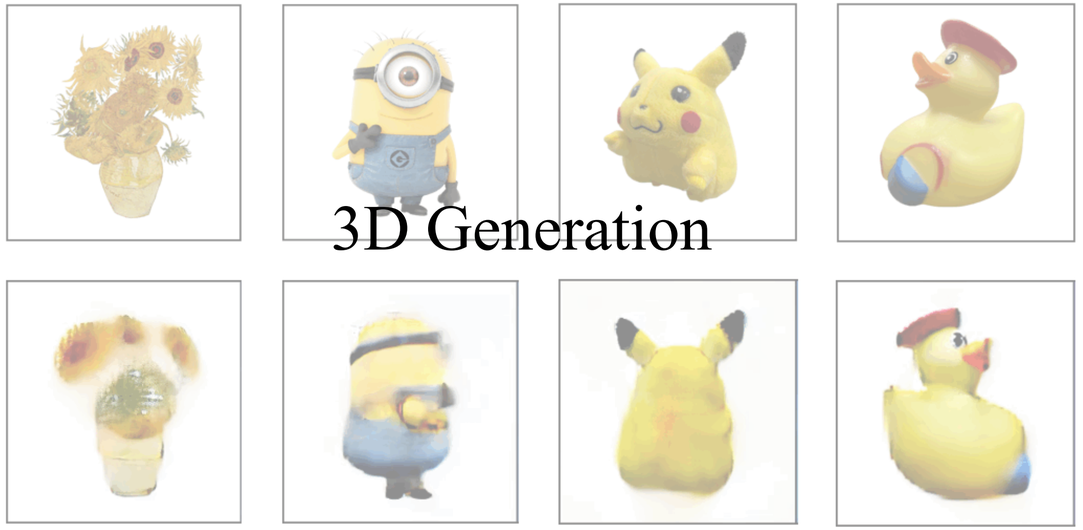
Homepage: https://www.sibowu.com/

Introduction



3D Reconstruction

?



3D Generation

Requires dense view coverage

Assumes single view input

Introduction

We propose **GenFusion:** A artifact-free 3D reconstruction and content expansion method

- utilise reconstruction-driven video generation model
- propose masked 3D reconstruction
- enable view interpolation and extrapolation tasks

Related Work

Regularization Techniques

- Unsupervised regularisation [1,2], Pseudo labels [3], Generative priors [4,5]
- Limitations: struggling with view extrapolation

Feed-forward 3D Reconstruction [6-8]

Limitations: limited to a small number of views

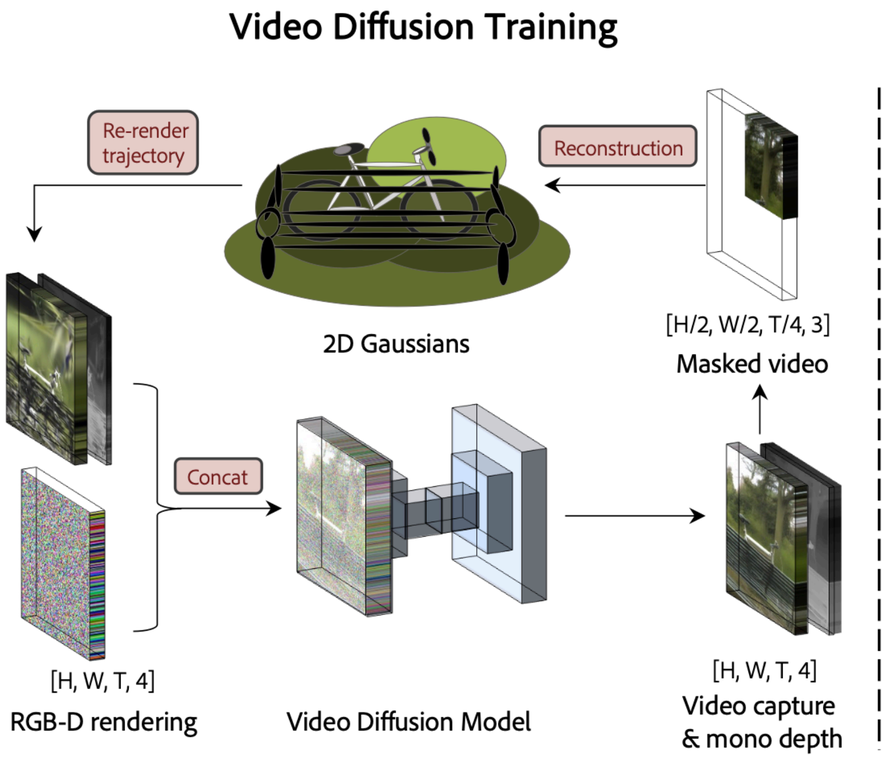
View-Conditioned Generation

- Image to image diffusion
- Multi view diffusion, e.g. CAT3D [9]
- Video diffusion, e.g. ViewCrafter [10]

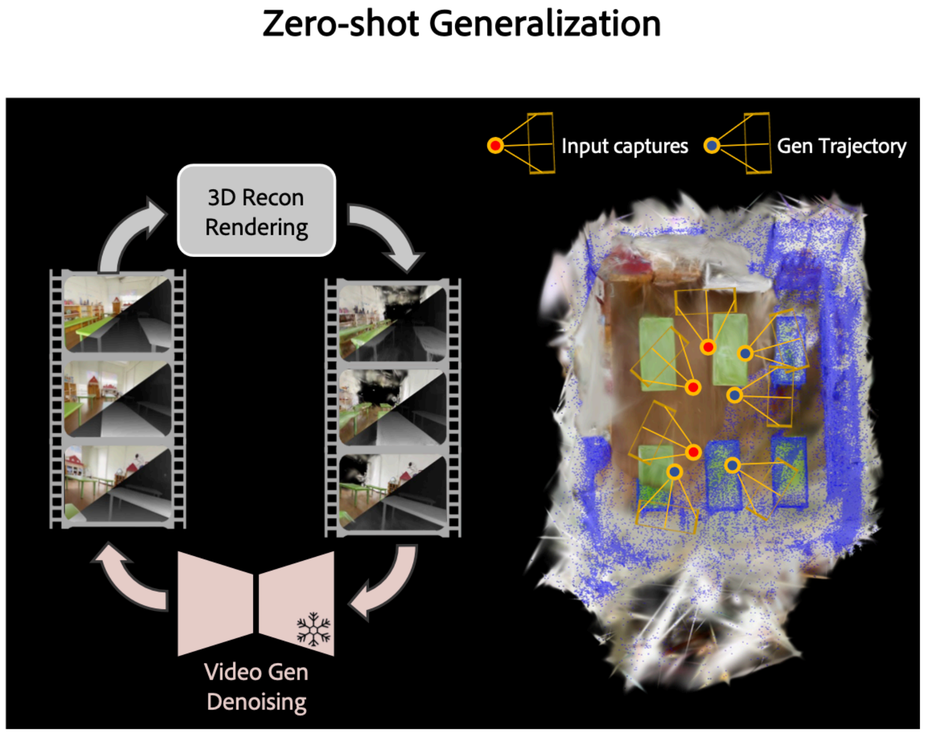
[1] Chen, Zheng, et al. "Structnerf: Neural radiance fields for indoor scenes with structural hints." *TPMI* (2023).
[2] Fu, Qiancheng, et al. "Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction." *NIPS (2022)*.
[3] Zhu, Zehao, et al. "Fsgs: Real-time few-shot view synthesis using gaussian splatting." *ECCV (2025)*.
[4] Liu, Xi, Chaoyi Zhou, and Siyu Huang. "3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors." *arXiv* (2024).
[5] Wu, Rundi, et al. "Reconfusion: 3d reconstruction with diffusion priors." CVPR (2024).
[6] Chen, Anpei, et al. "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo." *CVPR (2021)*
[7] Xu, Haofei, et al. "Murf: Multi-baseline radiance fields." CVPR (2024)
[8] Chen, Yuedong, et al. "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images." *ECCV (2025)*
[9] Gao, Ruiqi, et al. "Cat3d: Create anything in 3d with multi-view diffusion models." *arXiv* (2024).
[10] Yu, Wangbo, et al. "Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis." *arXiv* (2024).

Method

Key idea: align 3D reconstruction and generation through video renderings in cyclic approach



Stage 1: Train generation model to repair artefact renderings



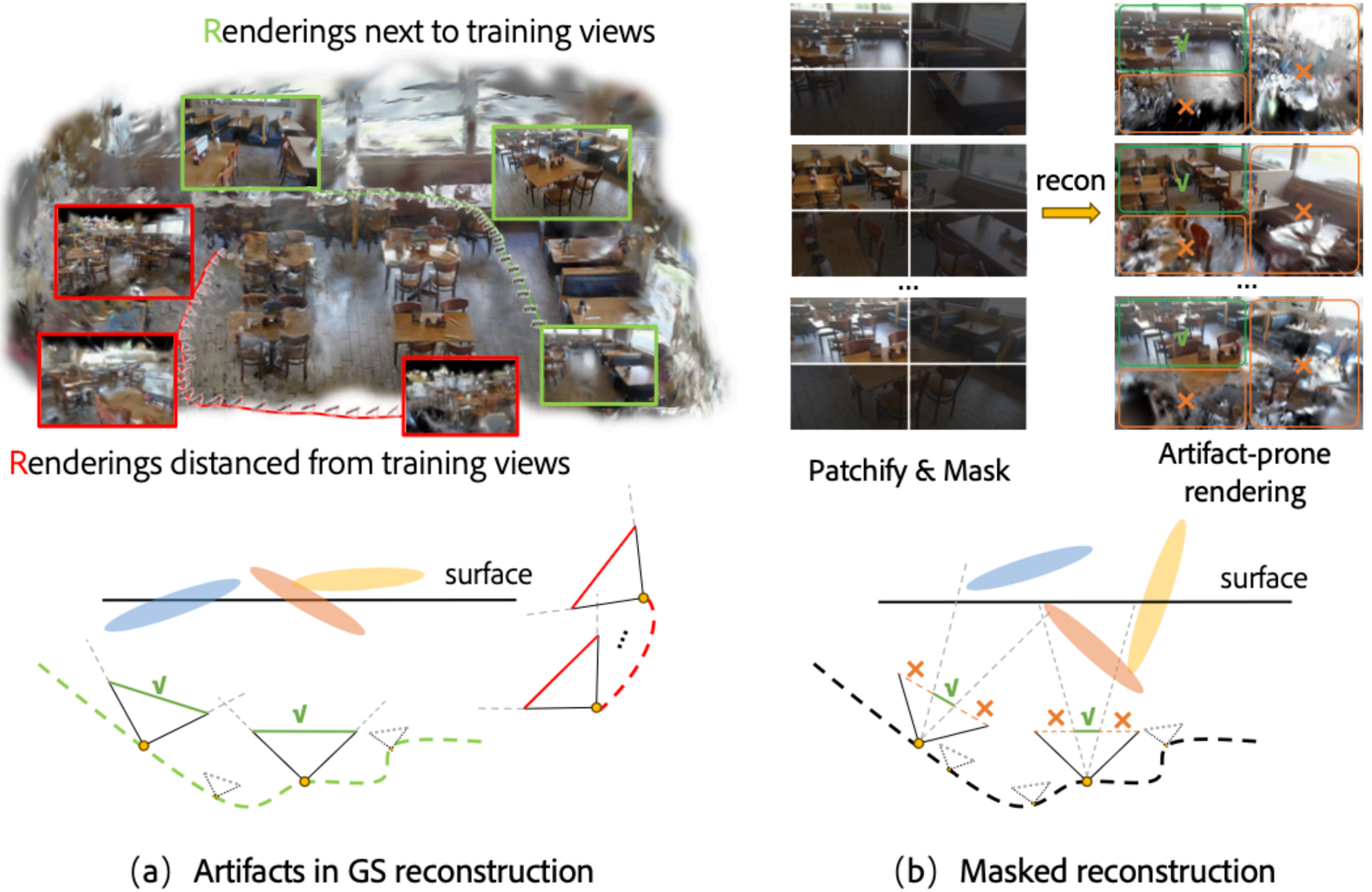
Stage 2: Optimize reconstruction method by leveraging pre-trained generative priors

Method

Stage 1: Reconstruction-driven Generation

Masked 3D reconstruction

- Simulates narrower field-of-view
- Train on masked inputs
- Test on full frames
- Note: mask is applied per scene, not per view



Method

Stage 1: Reconstruction-driven Generation

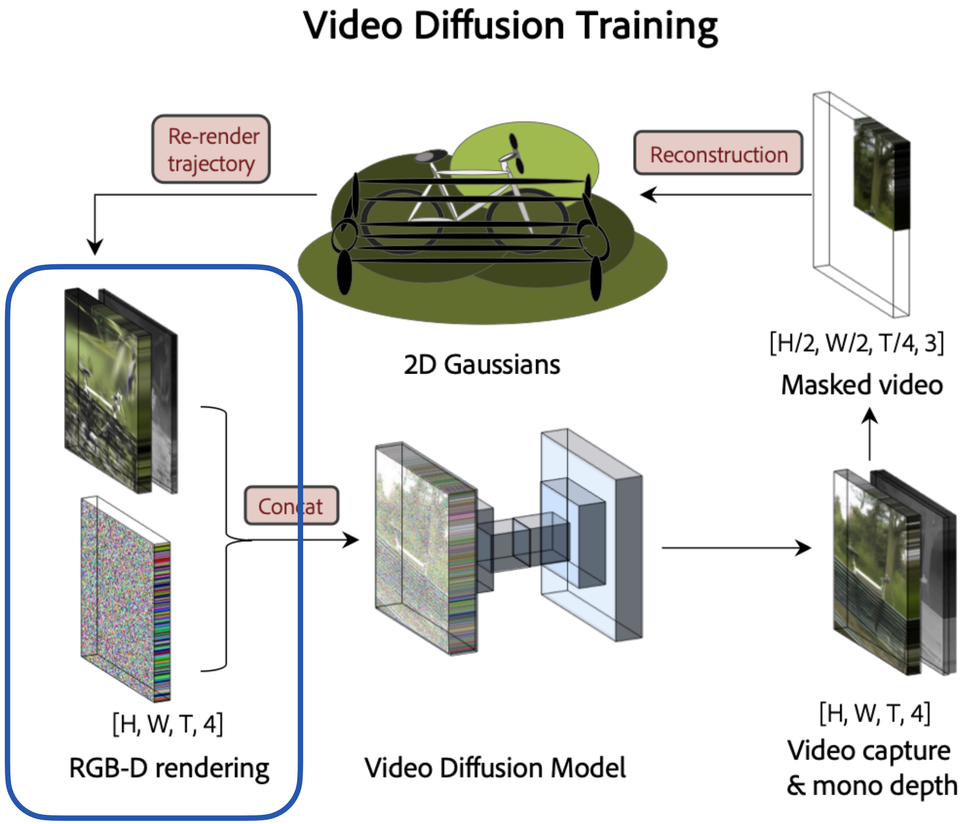
- Video Generation Model

  Frame consistency enhancement

  Multiple conditions

  Optimization goal:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right]$$

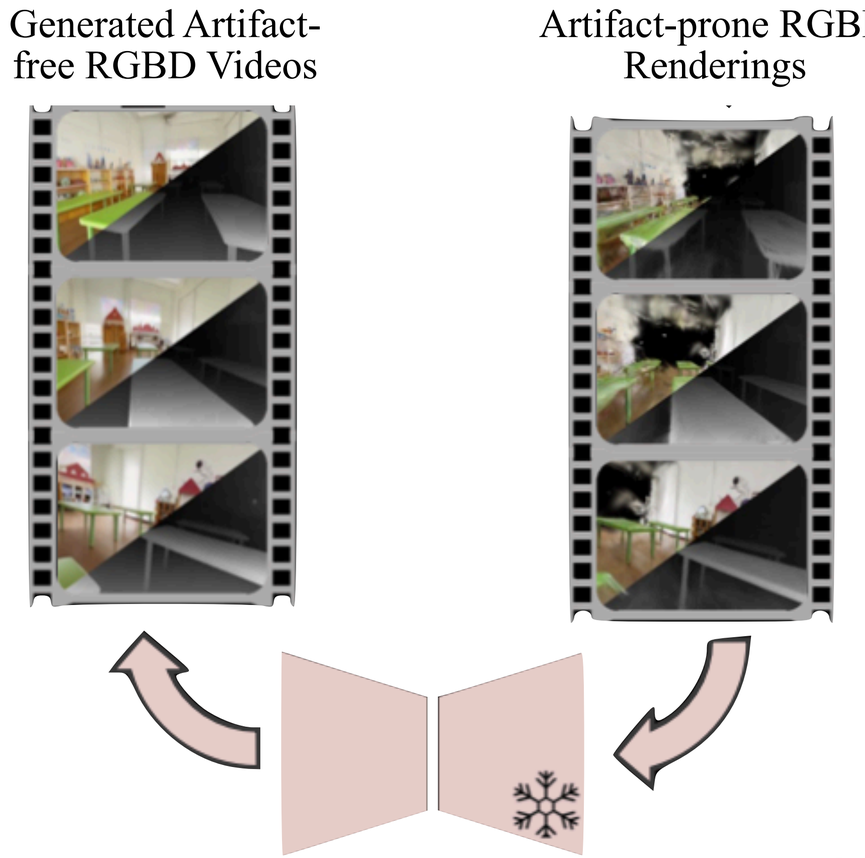


[11] Xing, Jinbo, et al. "Dynamicrafter: Animating open-domain images with video diffusion priors." ECCV (2025)

Method

Stage 2: Cyclic Fusion

Optimisation process: reconstruction and generation cycle



- Color supervision and depth supervision on both input views and generated views
- Regularisation weight: sinusoidal warm-up and annealing strategy

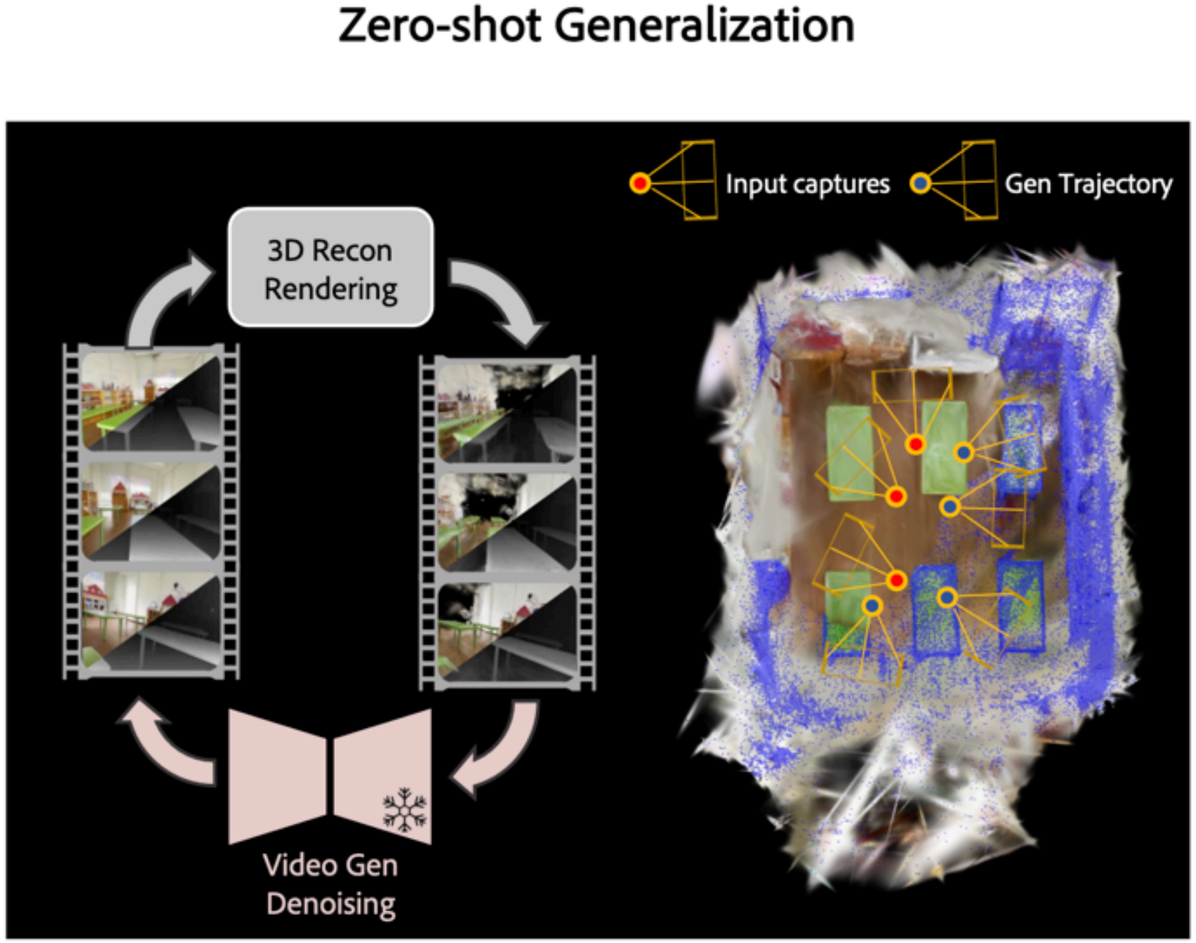
[12] Huang, Binbin, et al. "2d gaussian splatting for geometrically accurate radiance fields." SIGGRAPH (2024)

Method

Stage 2: Cyclic Fusion

artifact correction in under-observed regions

content creation for invisible areas



Method

Stage 2: Cyclic Fusion

Content expansion

Observation: Black holes in large unobserved area -> challenging to split and clone new Gaussians

Solution: adaptively adding new Gaussian points to the scene by unreliable map

Experiments & Results

Experimental Setup

Training set: DL3DV-10K [13]

- Optimise every scene for 7k iterations on data preparation stage
- Downsampled the number of training views to 1/4
- Resolution: 960x540

Evaluation set:

- DL3DV-Benchmark [13]: 24 scenes
- Tnt [14]: 7 scenes
- Mip-NeRF360 [15]: 9 scenes

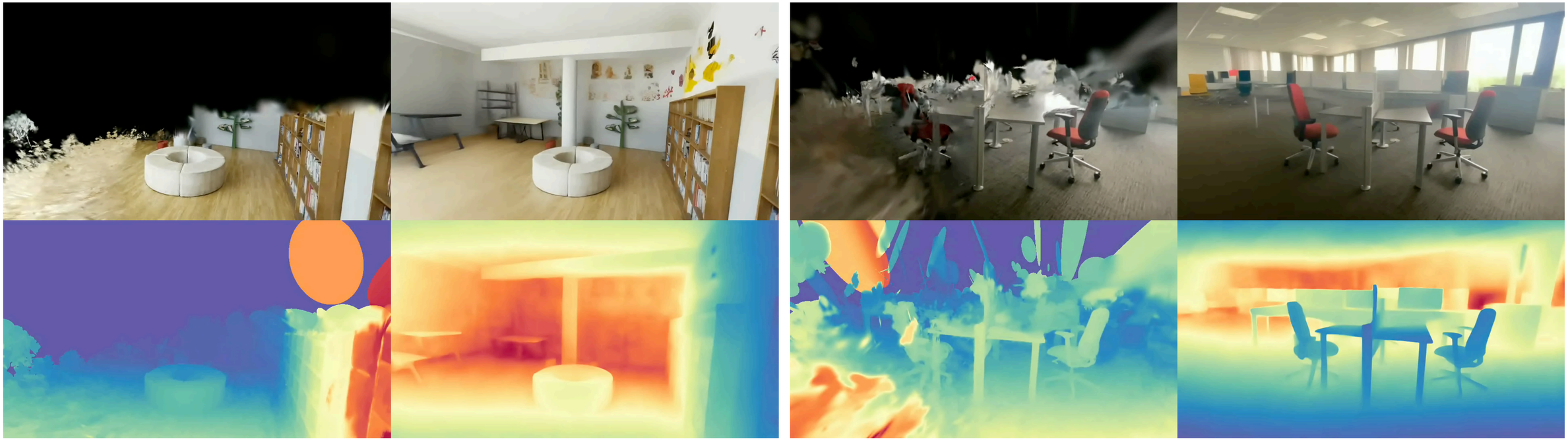
[13] Ling, Lu, et al. "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision." CVPR (2024)
[14] Knapitsch, Arno, et al. "Tanks and temples: Benchmarking large-scale scene reconstruction." TOG (2017)
[15] Barron, Jonathan T., et al. "Mip-nerf 360: Unbounded anti-aliased neural radiance fields." CVPR (2022)

Experiments & Results

Video Generation

Our video diffusion model is able to generate realistic RGB-D video from artifact- prone RGB-D renderings

Renderings | Generated | Renderings | Generated



Artifact-prone Renderings vs Generated RGB-D Videos

Experiments & Results

View Interpolation

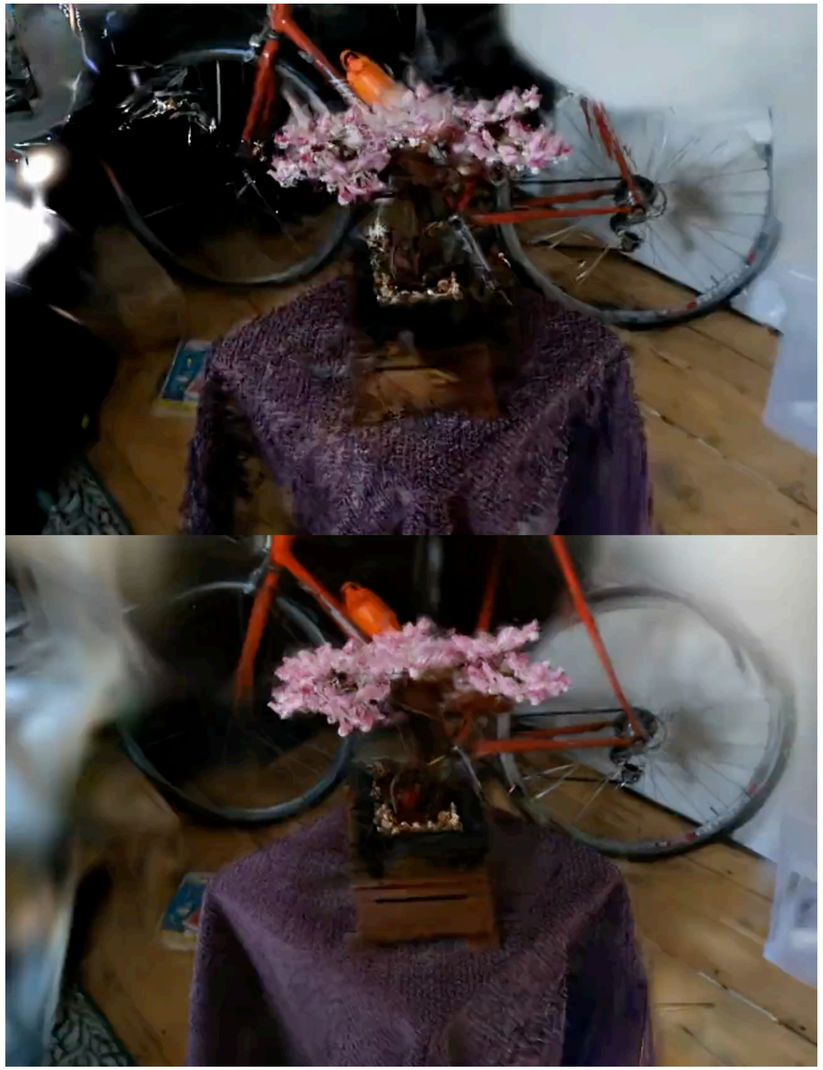
Qualitative comparison of novel view synthesis using sparse view input on Mip-NeRF360 scenes

Color: best, second best, third best

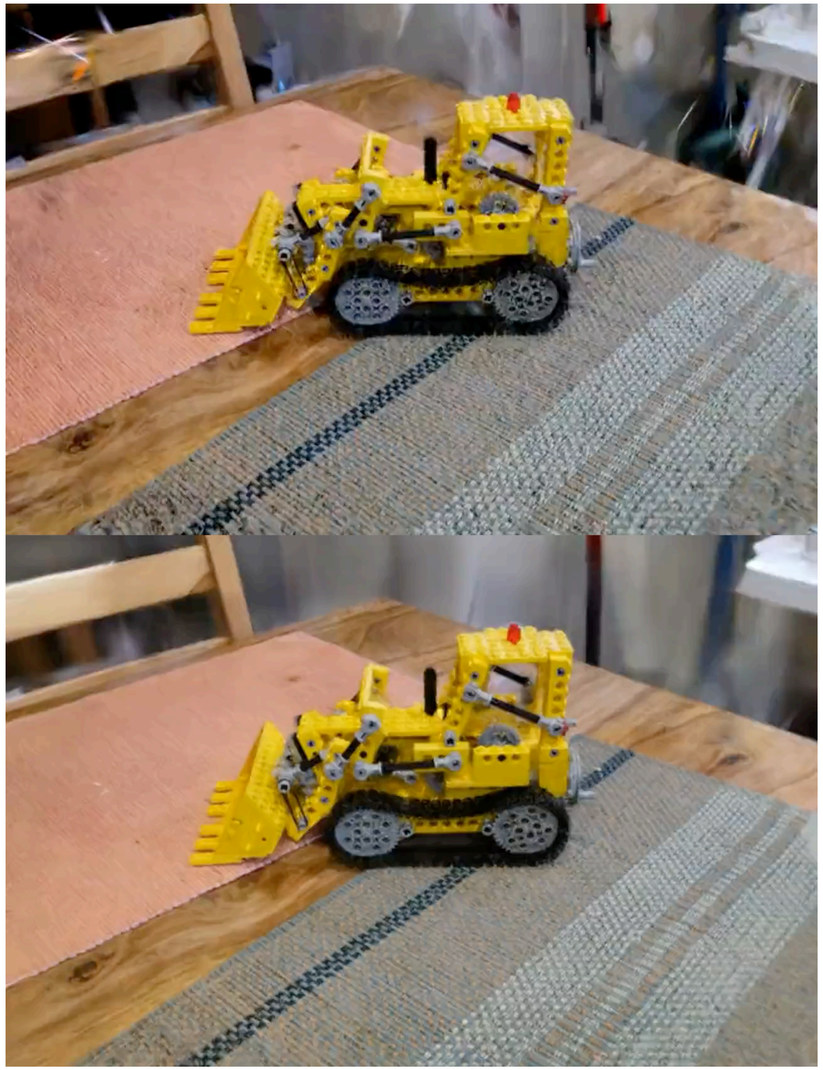
	PSNR ↑				SSIM ↑				LPIPS ↓			
	3-view	6-view	9-view	Avg.	3-view	6-view	9-view	Avg.	3-view	6-view	9-view	Avg.
Zip-NeRF [2]	12.77	13.61	14.30	13.56	0.271	0.284	0.312	0.289	0.705	0.663	0.633	0.667
DiffusioNeRF [47]	11.05	12.55	13.37	12.32	0.189	0.255	0.267	0.237	0.735	0.692	0.680	0.702
FreeNeRF [52]	12.87	13.35	14.59	13.60	0.260	0.283	0.319	0.287	0.715	0.717	0.695	0.709
SimpleNeRF [40]	13.27	13.67	15.15	14.03	0.283	0.312	0.354	0.316	0.741	0.721	0.676	0.713
ZeroNVS [39]	14.44	15.51	15.99	15.31	0.316	0.337	0.350	0.334	0.680	0.663	0.655	0.666
ReconFusion [46]	15.50	16.93	18.19	16.87	0.358	0.401	0.432	0.397	0.585	0.544	0.511	0.547
3DGS [25]	13.06	14.96	16.79	14.94	0.251	0.355	0.447	0.351	0.576	0.505	0.446	0.509
2DGS [23]	13.07	15.02	16.67	14.92	0.243	0.338	0.423	0.335	0.580	0.506	0.449	0.512
FSGS [59]	14.17	16.12	17.94	16.08	0.318	0.415	0.492	0.408	0.578	0.517	0.468	0.521
GenFusion (Ours)	15.29	17.16	18.36	16.93	0.369	0.447	0.496	0.437	0.585	0.500	0.465	0.517

Quantitative evaluation of sparse view 3D reconstruction methods on Mip-NeRF360 dataset.

3 views



6 views



9 views

Qualitative comparison of novel view synthesis using sparse view input on Mip-NeRF360 scenes
(Up: 2DGS, Down: Ours)

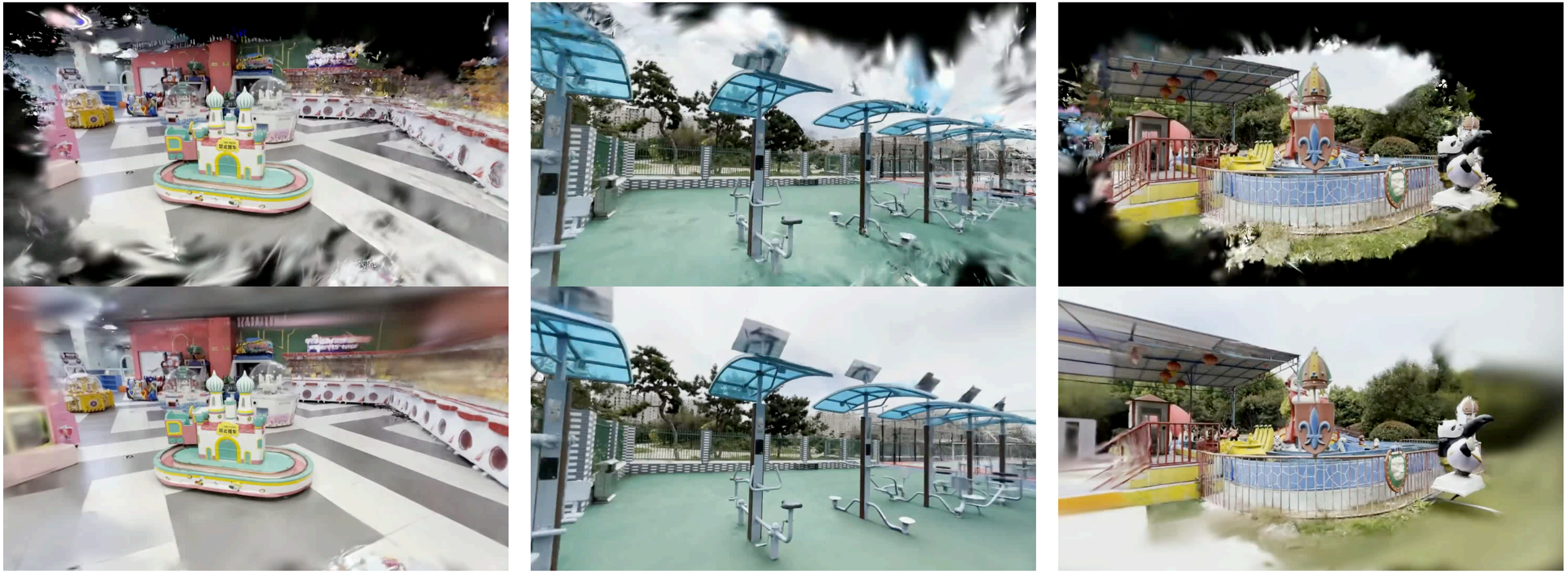
Experiments & Results

View Extrapolation

Quantitative comparison of novel view synthesis using masked input on DL3DV scenes

Color: best, second best, third best

	DL3DV 1/2 fps.			TnT 1/2 fps.			DL3DV 1/4 fps.			TnT 1/4 fps.		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS [25]	17.22	0.740	0.314	15.95	0.653	0.414	16.90	0.728	0.321	14.75	0.609	0.440
2DGS [23]	16.56	0.717	0.323	15.46	0.640	0.409	16.02	0.693	0.336	14.38	0.589	0.440
FSGS [59]	18.25	0.722	0.362	16.72	0.625	0.465	17.83	0.710	0.370	16.04	0.607	0.473
Ours	20.47	0.788	0.284	17.45	0.662	0.427	20.01	0.780	0.292	16.29	0.630	0.447



Qualitative comparison of novel view synthesis using masked input on DL3DVscenes (Up: 2DGS, Down: Ours)

Experiments & Results

Scene Completion

Discussion

Summary

We propose GenFusion: leveraging cyclical reconstruction and generation loop for view interpolation and extrapolation

Efficiently close the loop between 3D reconstruction and generation

Limitations

Optimisation time increases (ab. 40min)

Blurriness caused by inconsistency between frames segment

References

[1] Chen, Zheng, et al. "Structnerf: Neural radiance fields for indoor scenes with structural hints." *TPMI* (2023).
[2] Fu, Qiancheng, et al. "Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction." *NIPS (2022)*.
[3] Zhu, Zehao, et al. "Fsgs: Real-time few-shot view synthesis using gaussian splatting." *ECCV (2025)*.
[4] Liu, Xi, Chaoyi Zhou, and Siyu Huang. "3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors." *arXiv* (2024).
[5] Wu, Rundi, et al. "Reconfusion: 3d reconstruction with diffusion priors." CVPR (2024).
[6] Chen, Anpei, et al. "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo." *CVPR (2021)*
[7] Xu, Haofei, et al. "Murf: Multi-baseline radiance fields." CVPR (2024)
[8] Chen, Yuedong, et al. "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images." *ECCV (2025)*
[9] Gao, Ruiqi, et al. "Cat3d: Create anything in 3d with multi-view diffusion models." *arXiv* (2024).
[10] Yu, Wangbo, et al. "Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis." *arXiv* (2024).
[11] Xing, Jinbo, et al. "Dynamicrafter: Animating open-domain images with video diffusion priors." ECCV (2025)
[12] Huang, Binbin, et al. "2d gaussian splatting for geometrically accurate radiance fields." SIGGRAPH (2024)a
[13] Ling, Lu, et al. "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision." CVPR (2024)
[14] Knapitsch, Arno, et al. "Tanks and temples: Benchmarking large-scale scene reconstruction." TOG (2017)
[15] Barron, Jonathan T., et al. "Mip-nerf 360: Unbounded anti-aliased neural radiance fields." CVPR (2022)