# SGAM: Building a Virtual 3D World through Simultaneous Generation and Mapping

Yuan Shen, Wei-Chiu Ma, Shenlong Wang

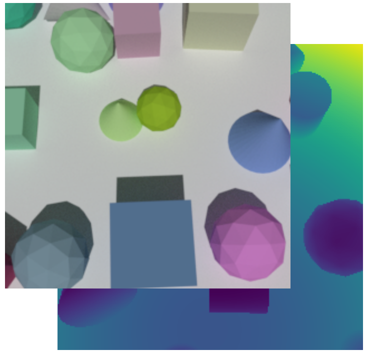


UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Massachusetts Institute of Technology

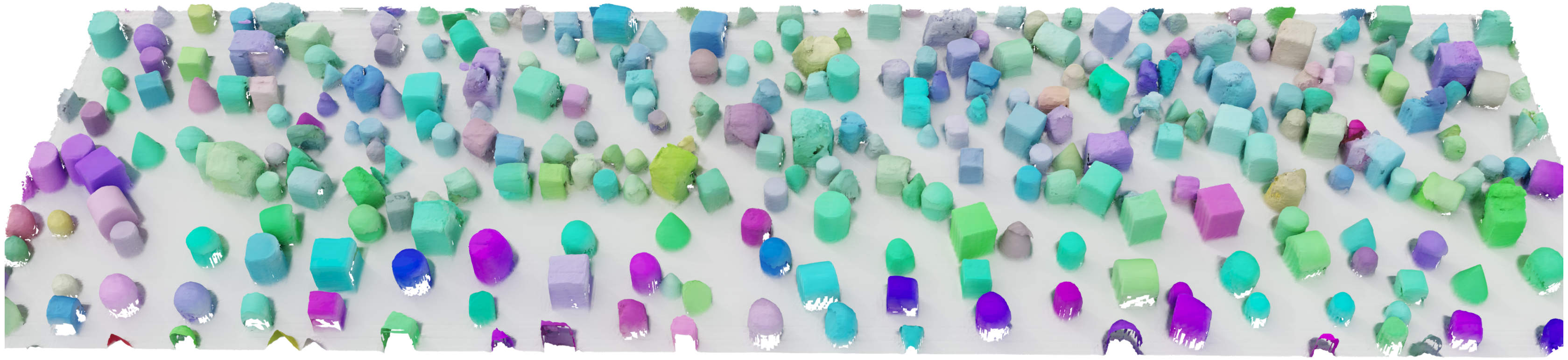
# Goal

# Goal



Input: a RGB-D image

# Goal



Input: a RGB-D image

Output: a consistent, realistic and large-scale 3D world.

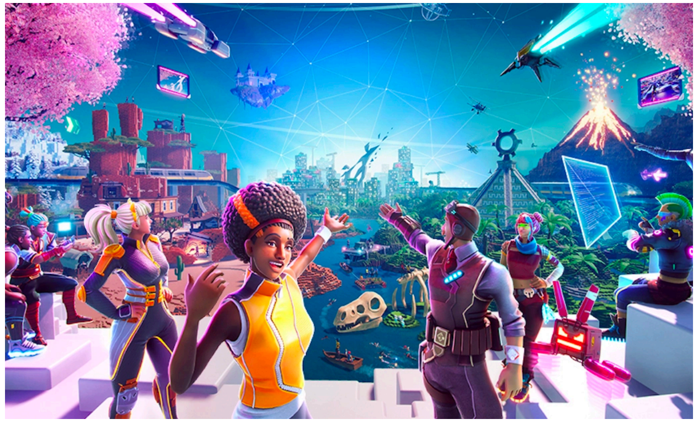
# Applications for 3D Generation

Gaming

Simulation

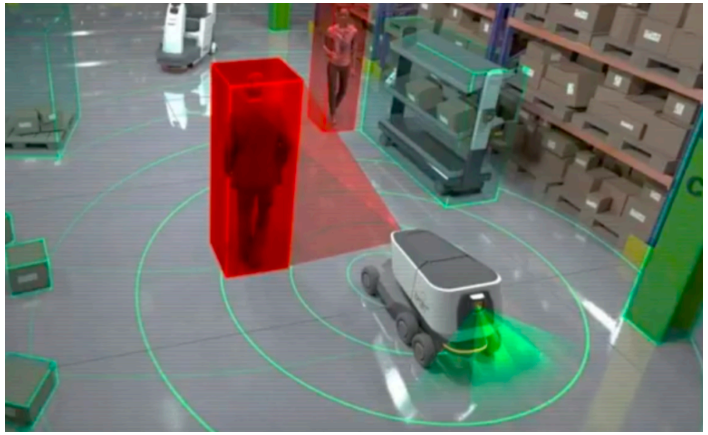
Filmmaking


Metaverse



Urban Planning

Robot Navigation

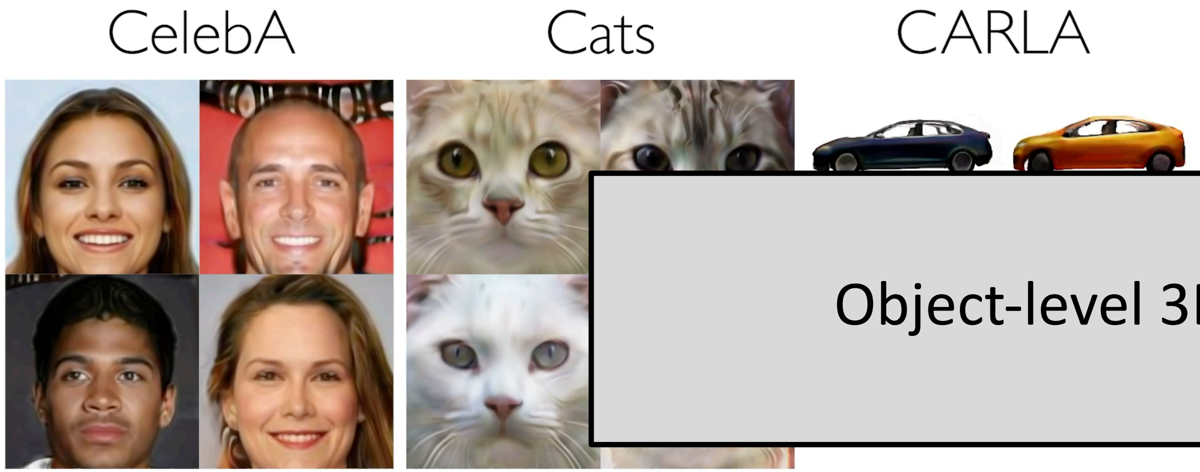


# Desiderata

- Scalability
- Realism
- Consistency

Source: Google Earth

# 3D Object Generation



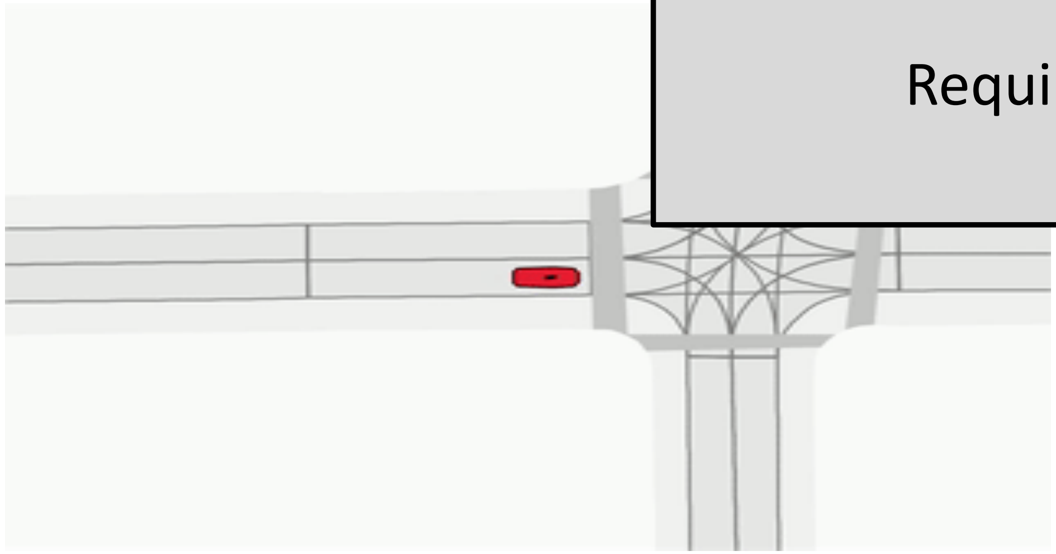

Object-level 3D generation

Chan et al. pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis, CVPR 2021

Gao et al. GET3D: A Generative Model of High Quality 3D Textured Shapes Learned from Images, NeurIPS 2022

# 3D World Generation

Devaranjan* and Kar* et al., ECCV 2020.



Tan et al., CVPR 2021.

Paschalidou et al., NeurIPS 2021.

Requires additional assets

# Perpetual Video Generation

Do not produce a 3D world



Liu* and Tucker* et al., ICCV 2021.
Rombach* and Esser* et al, ICCV 2021.

# Texture synthesis

For each step:



Efros and Freeman, SIGGRAPH 2001.

# Texture synthesis

For each step:

- Grow the target image by selecting a new patch to fill



Efros and Freeman, SIGGRAPH 2001.

# Texture synthesis

For each step:

- Grow the target image by selecting a new patch to fill
- Sample a patch that is *consistent* with the current target image and is *realistic*

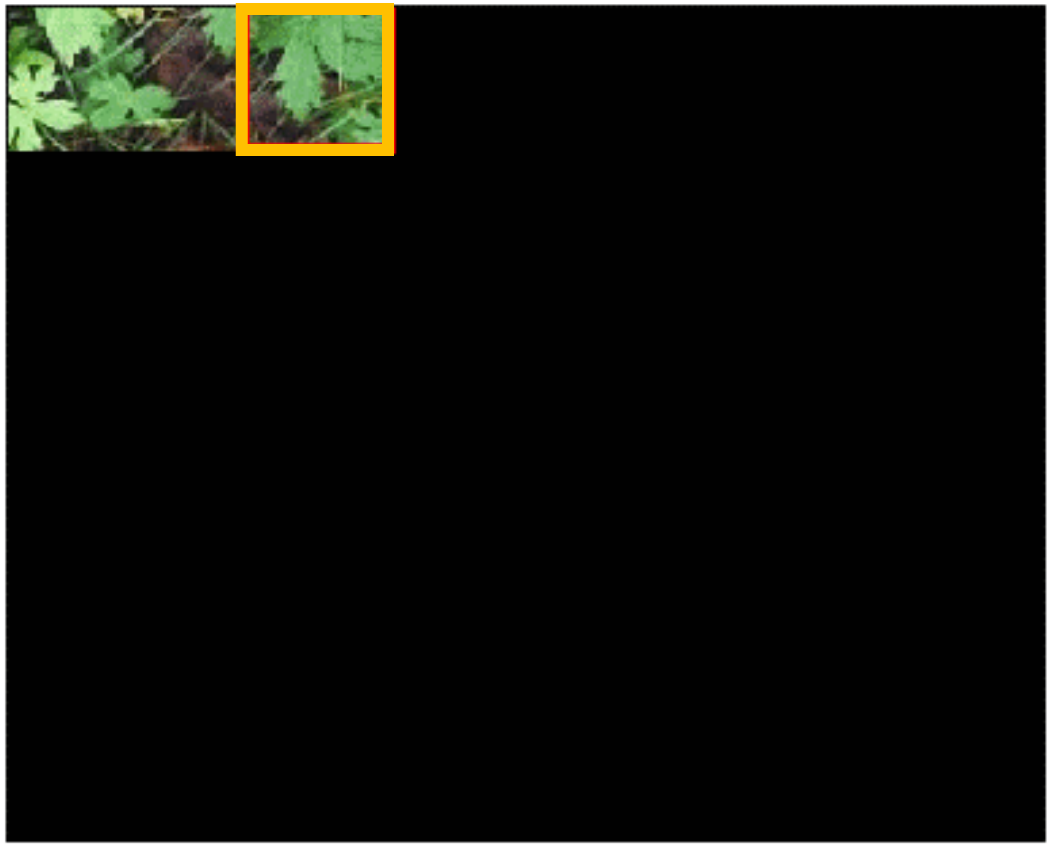


Efros and Freeman, SIGGRAPH 2001.

# Texture synthesis

For each step:

- Grow the target image by selecting a new patch location to fill
- Sample a patch that is *consistent* with the current target image and is *realistic*
- Add to the new image





Efros and Freeman, SIGGRAPH 2001.

# Texture synthesis

For each step:

- Grow the target image by selecting a new patch to fill
- Sample a patch that is *consistent* with the current target image and is *realistic*
- Add to the new image



Efros and Freeman, SIGGRAPH 2001.

# Texture synthesis

“2D world”

“Ortho camera at position (u, v) ”

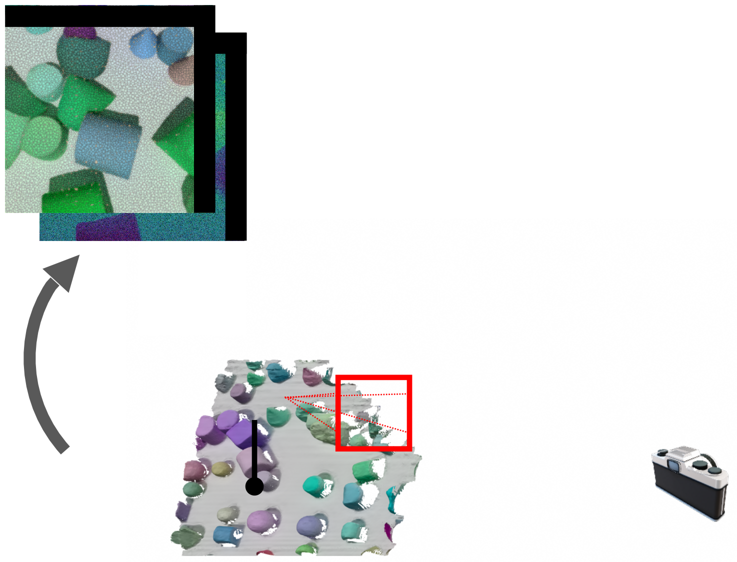


Efros and Freeman, SIGGRAPH 2001.

# Simultaneous Generation and Mapping

For each step:

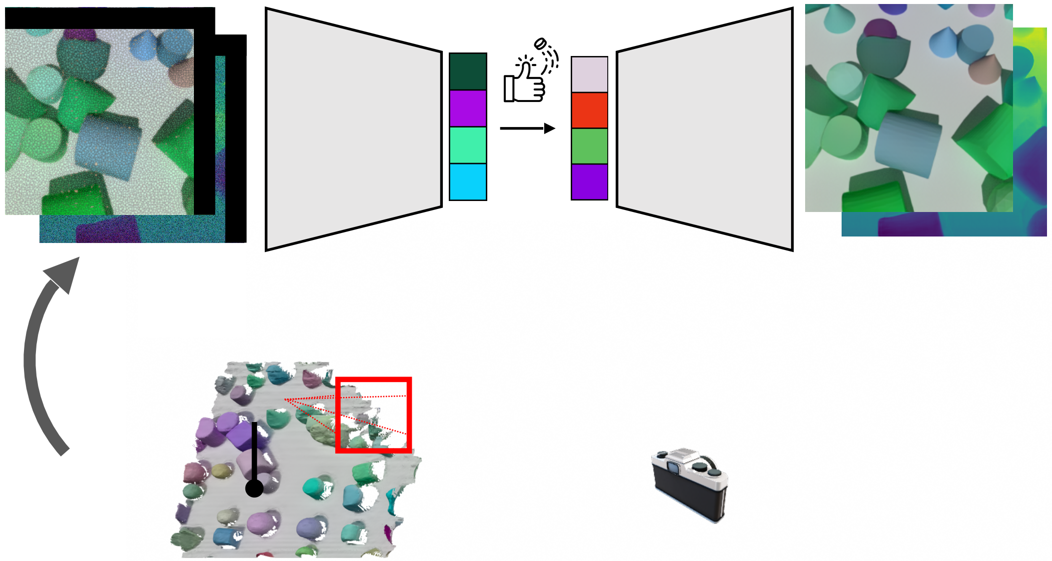
- Grow the target scene by selecting a view of interest



# Simultaneous Generation and Mapping

For each step:

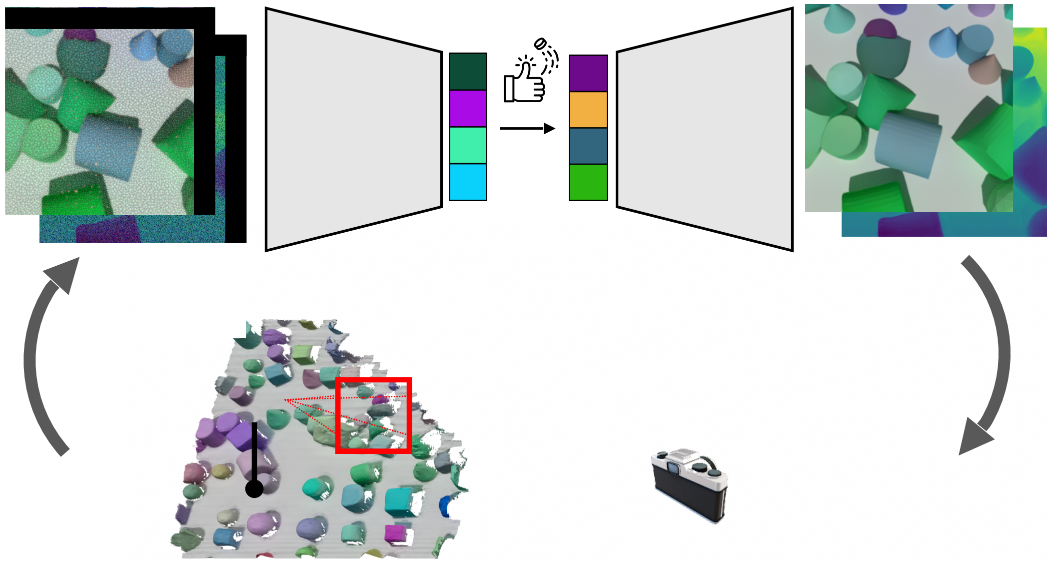
- Grow the target scene by selecting a view of interest
- Sample a new image that is *consistent* with the current 3D scene and is *realistic*



# Simultaneous Generation and Mapping

For each step:

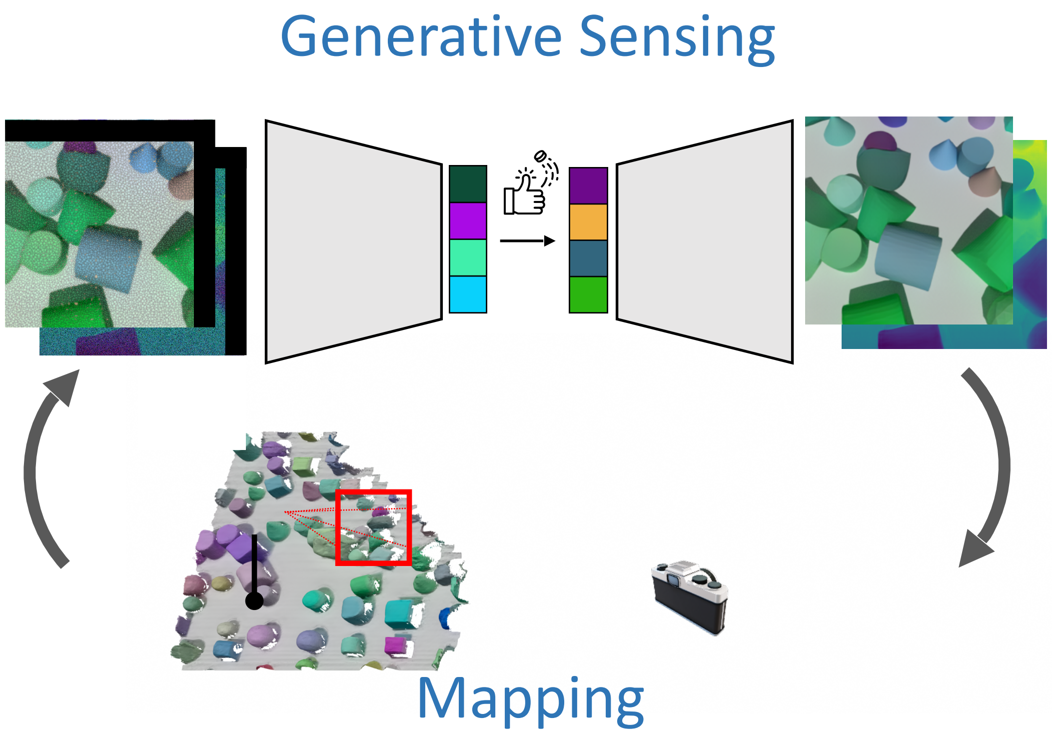
- Grow the target scene by selecting a view of interest
- Sample a new image that is *consistent* with the current 3D scene and is *realistic*
- Integrate it with the scene



# Simultaneous Generation and Mapping

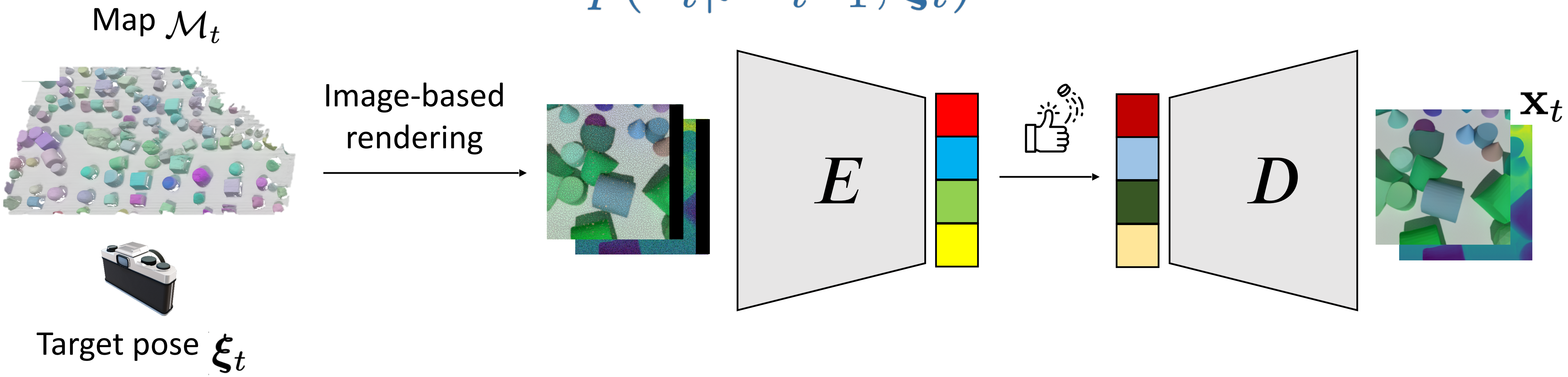
For each step:

- Grow the target scene by selecting a view of interest
- Sample a new image that is *consistent* with the current 3D scene and is *realistic*
- Integrate it with the scene



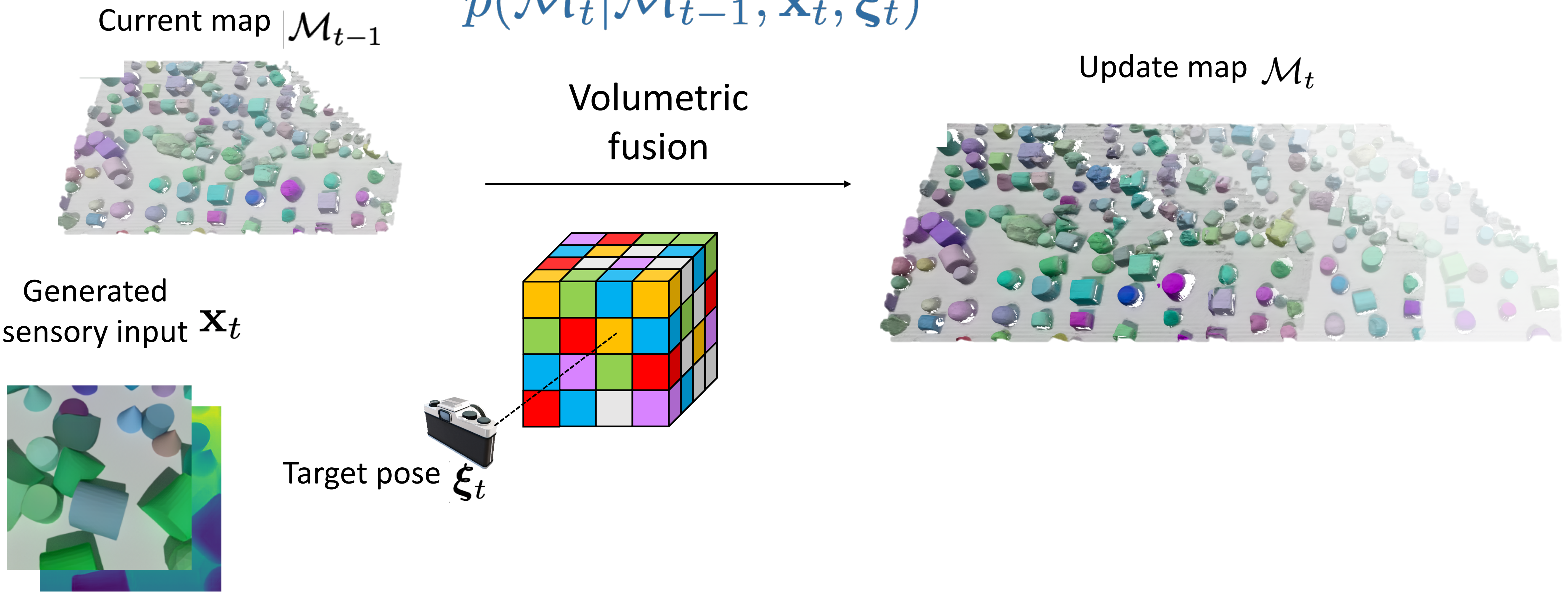
# Generative Sensing

$p(\mathbf{x}_t | \mathcal{M}_{t-1}, \boldsymbol{\xi}_t)$



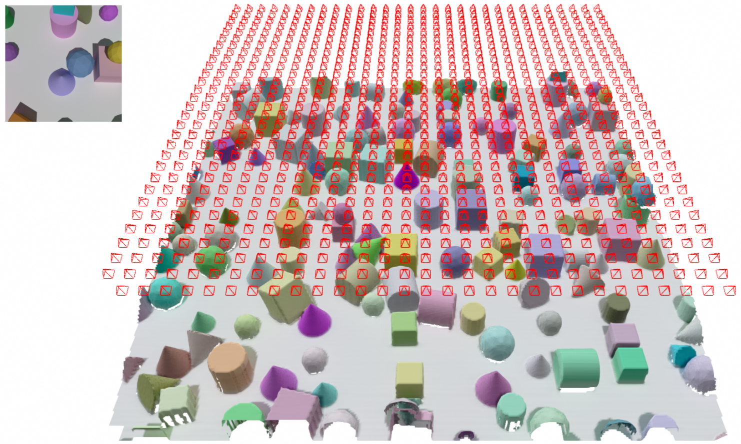
# Mapping

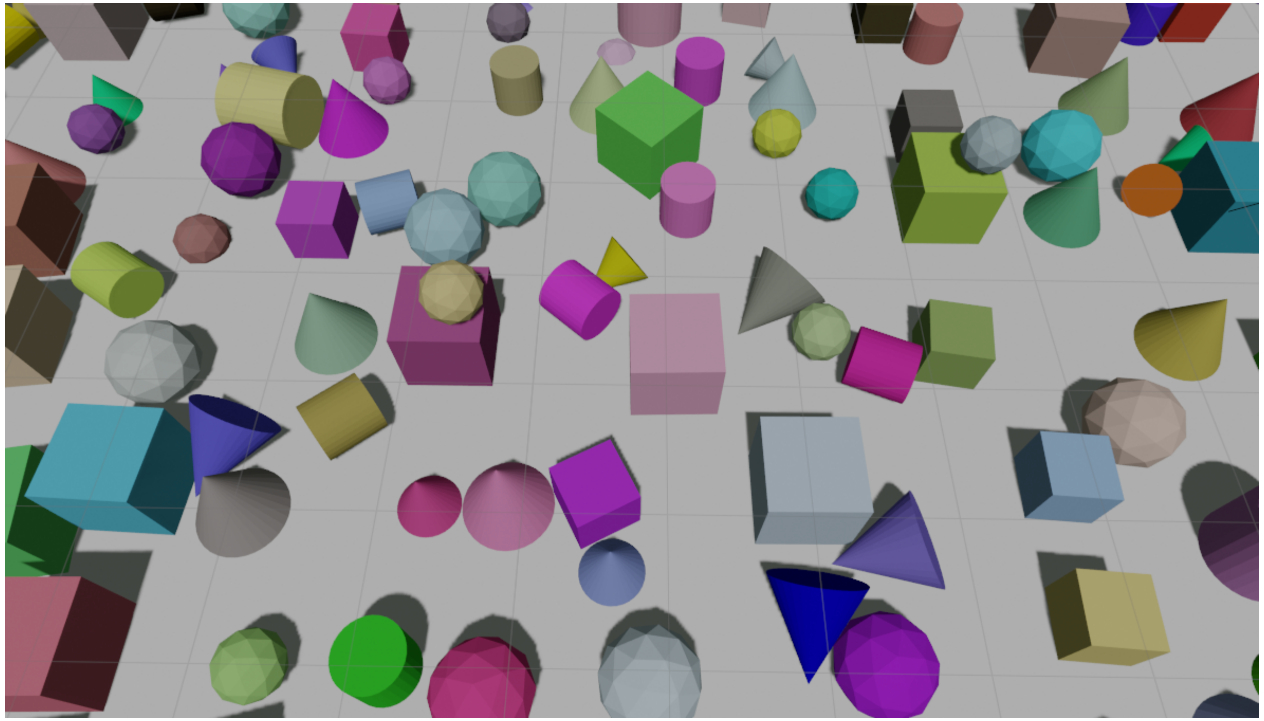
$$p(\mathcal{M}_t | \mathcal{M}_{t-1}, \mathbf{x}_t, \boldsymbol{\xi}_t)$$



# Dataset

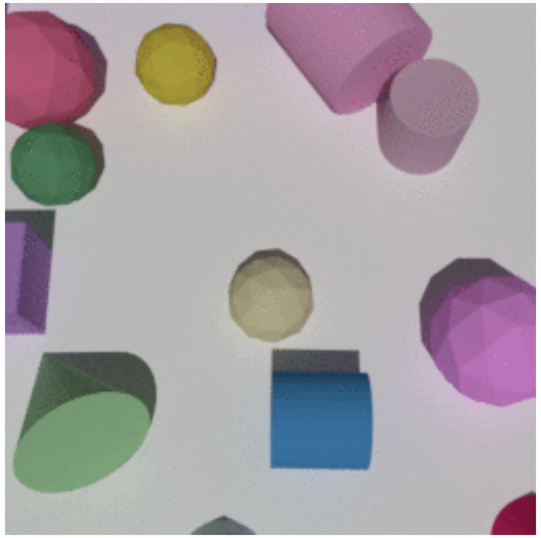
- CLEVR-Infinite: posed RGB-D synthetic data
- Google Earth: posed RGB-D real-world data





# Comparison on CLVER-Infinite

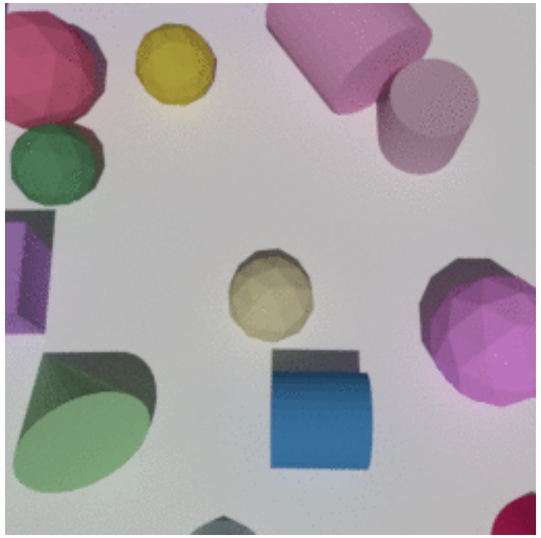
InfiniteNature



mode-collapse

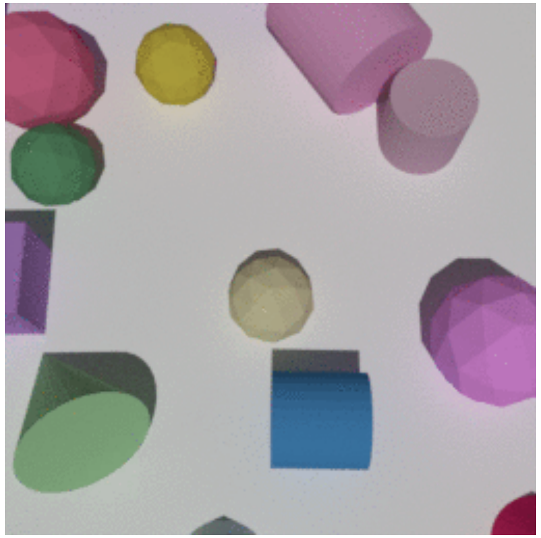
# Comparison on CLVER-Infinite

InfiniteNature



mode-collapse

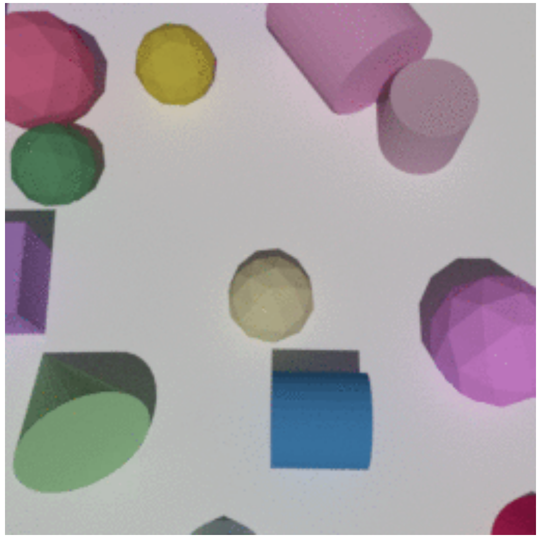
GFVS-Explicit



slow and early decoding issues

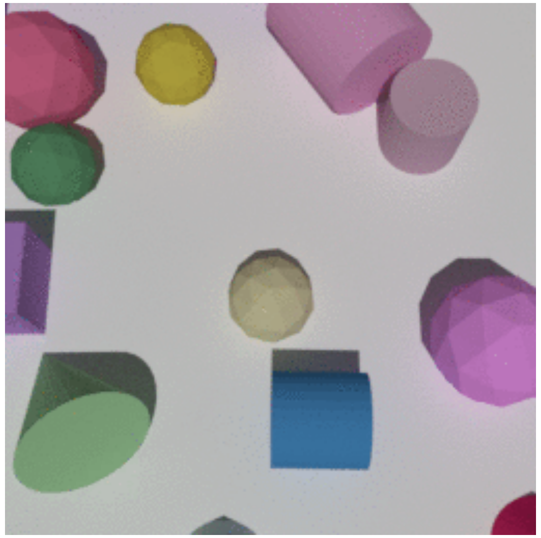
# Comparison on CLVER-Infinite

InfiniteNature



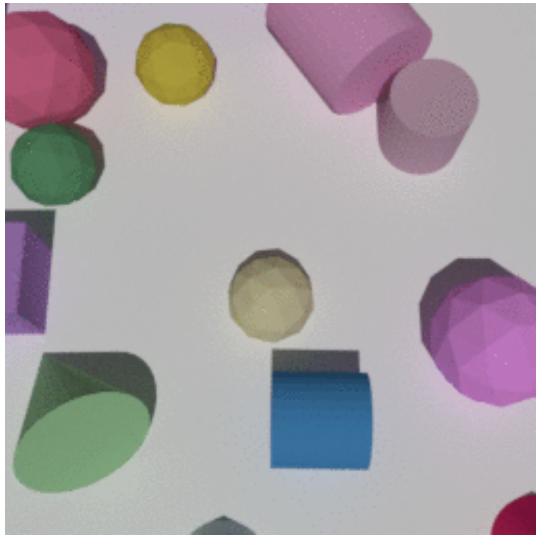
mode-collapse

GFVS-Explicit



slow and early decoding issues

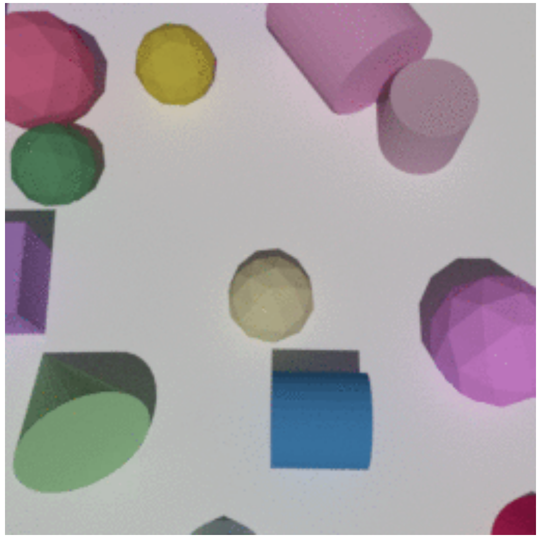
GFVS-Implicit



slow and inconsistent

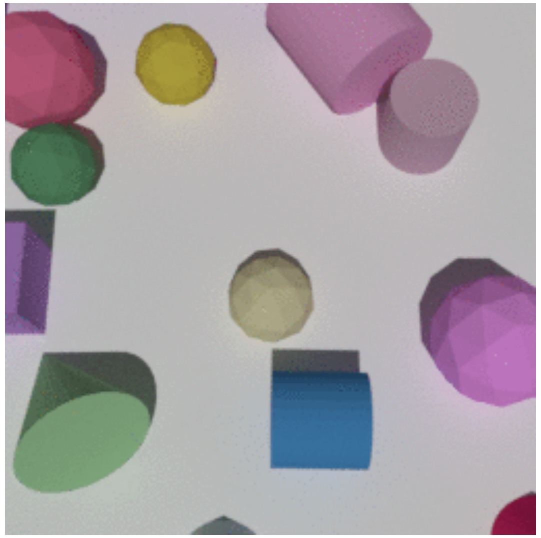
# Comparison on CLVER-Infinite

InfiniteNature



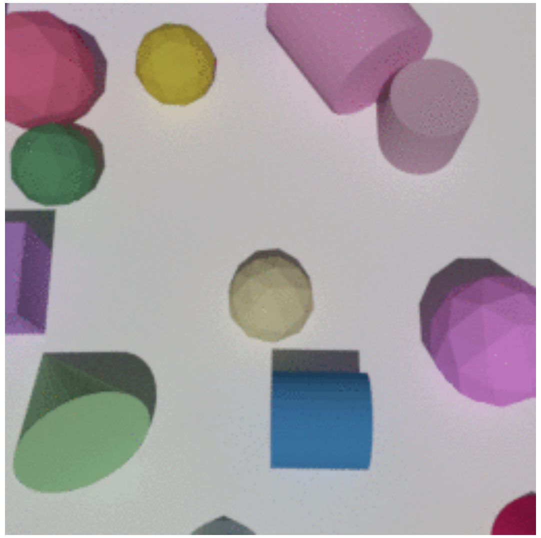
mode-collapse

GFVS-Explicit



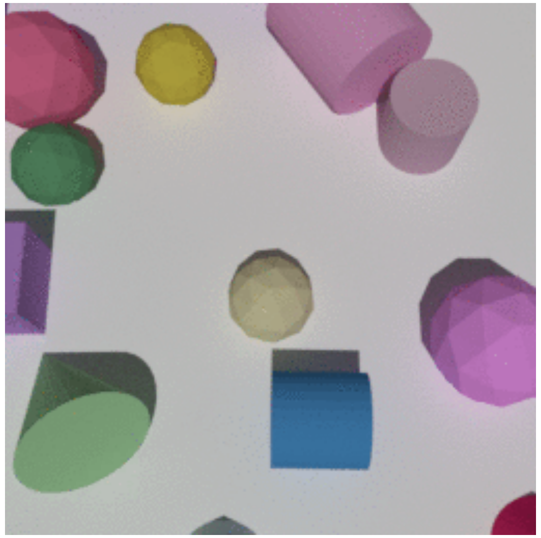
slow and early decoding issues

GFVS-Implicit



slow and inconsistent

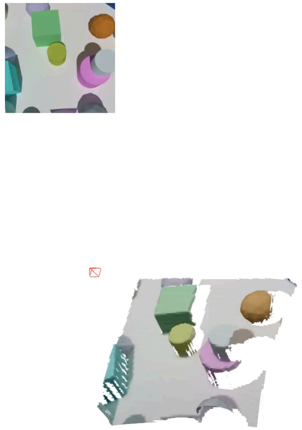
**Ours**



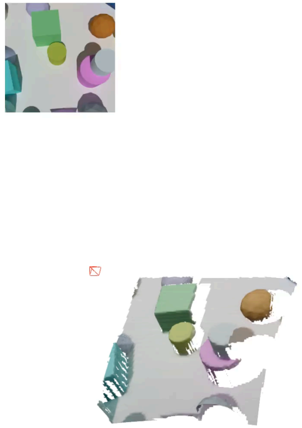
fast and high-quality

# Effect of different camera trajectories

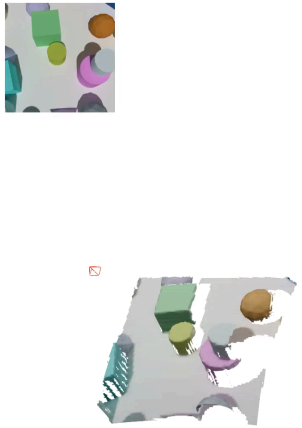
zig-zag



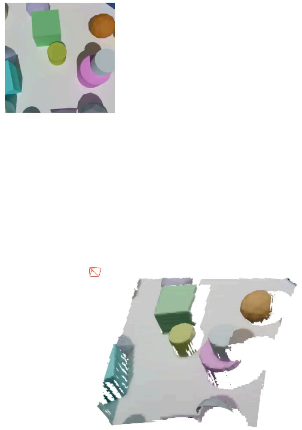
row-major



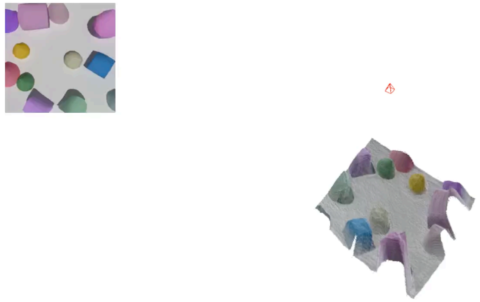
column-major



greedy

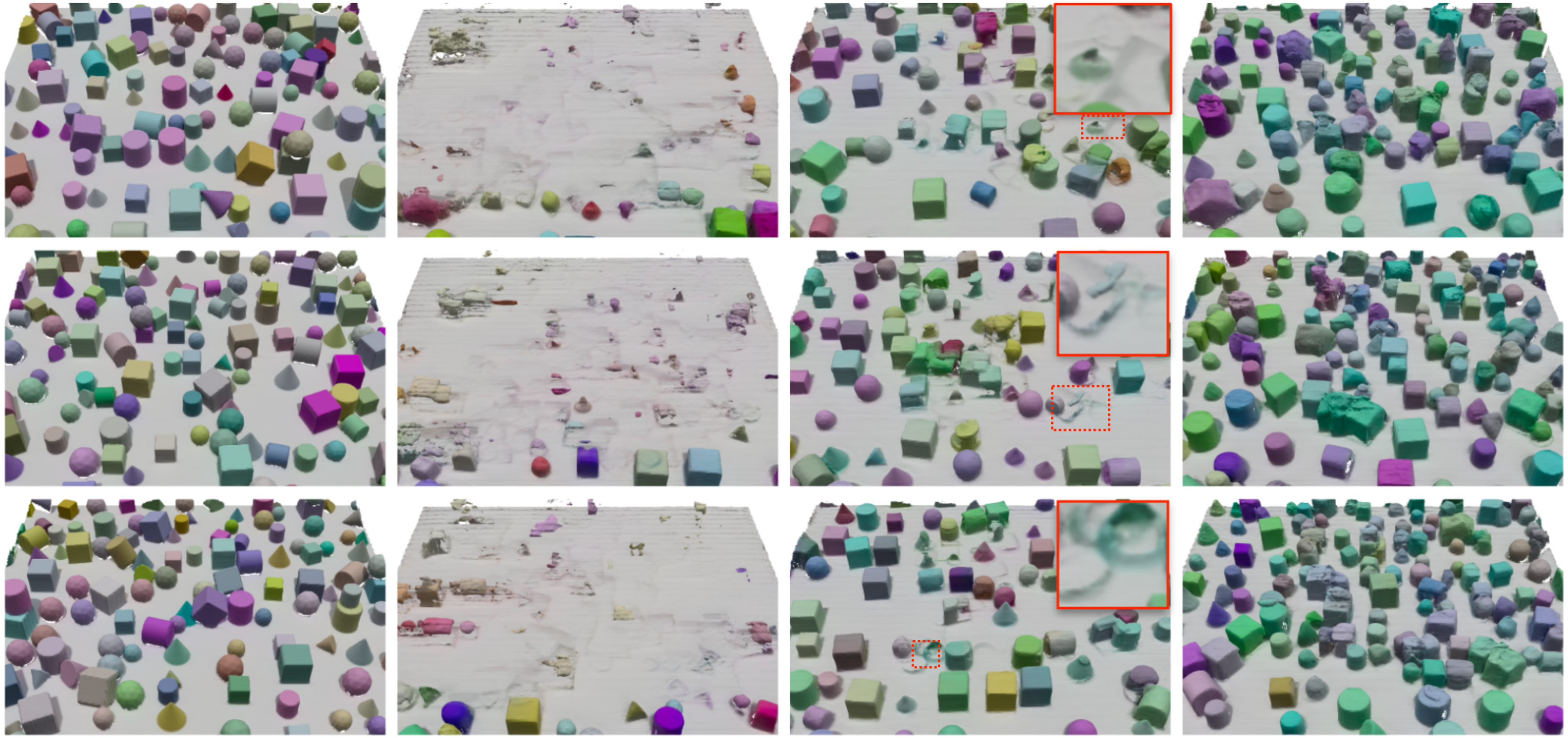


spiral



ring

# Scene-level Results on CLEVR-Infinite



GT | GFVS-Explicit [61] | GFVS-Implicit [61] | Ours

1. Robin Rombach & Patrick Esser, et al., “Geometry-Free View Synthesis: Transformers and no 3D Priors”, ICCV 2021

# Scene-level Results on CLEVR-Infinite

| renderer | FID ↓ | JSD ($10^{-2}$) ↓ | MMD ($10^{-5}$) ↓ |
| --- | --- | --- | --- |
| GFVS-implicit | **16.14** | 0.775 | 4.510 |
| GFVS-explicit | 82.82 | 10.870 | 211.500 |
| ours | 26.60 | **0.656** | **4.441** |

# Comparison on Google Earth

InfiniteNature

GFVS-Explicit

GFVS-Implicit

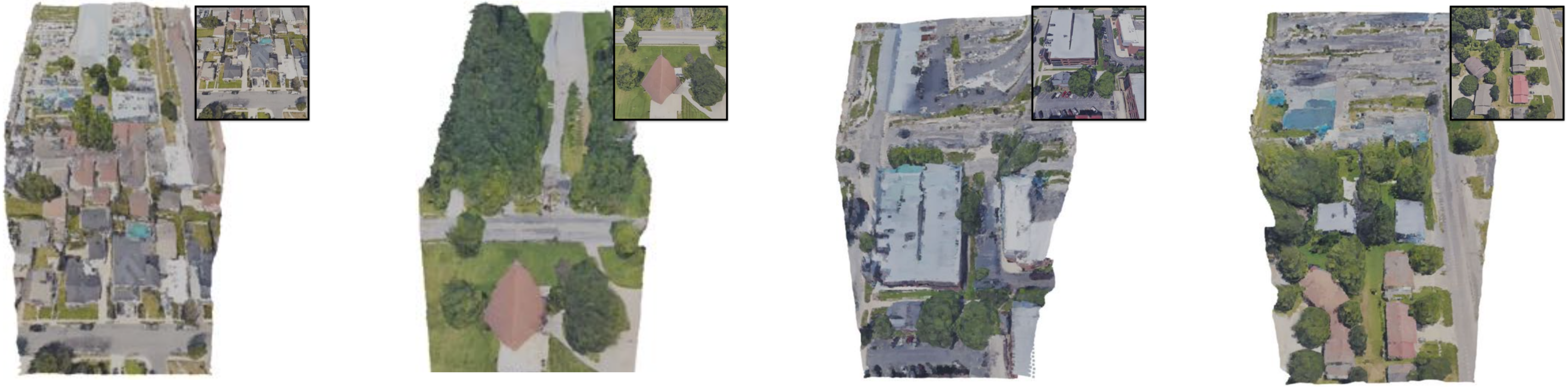
**Ours**



Liu* and Tucker* et al, ICCV 2021.
Rombach* and Esser* et al, ICCV 2021.

# More Qualitative Results on Google Earth

# Qualitative Results on Google Earth



# Scene-Level Results on Google Earth

| renderer | FID↓ |
|---|---|
| InfiniteNature | 182.6 |
| GFVS-implicit | 160.40 |
| GFVS-explicit | 113.12 |
| ours | **79.26** |

1. Andrew Liu & Richard Tucker, et al., *“Infinite Nature: Perceptual View Generation of Natural Scenes from a Single Image”*, ICCV 2021
2. Robin Rombach & Patrick Esser, et al., “Geometry-Free View Synthesis: Transformers and no 3D Priors”, ICCV 2021

# Conclusion

- We present SGAM — a 3D scene generation framework that produces realistic, consistent and large-scale 3D virtual world through **simultaneous generation and mapping**

- Take-home message: explicit 3D mapping helps consistency, realism and scalability for perpetual generation.

Thanks for watching, please check our paper for details!

Project Page: https://yshen47.github.io/sgam/

Codebase link: https://github.com/yshen47/SGAM

Paper link: https://openreview.net/forum?id=17KCLTbRymw

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Massachusetts Institute of Technology