Boost Perception Model in Autonomous Driving by Generative AI

Leheng Li 李乐恒

Ph.D. student at HKUST(GZ)

14 August 2023

Contents

- Basic of NeRF
- Recent work of NeRF in autonomous driving
- Generative NeRF helps downstream task (Lift3D)

Background of Leheng Li

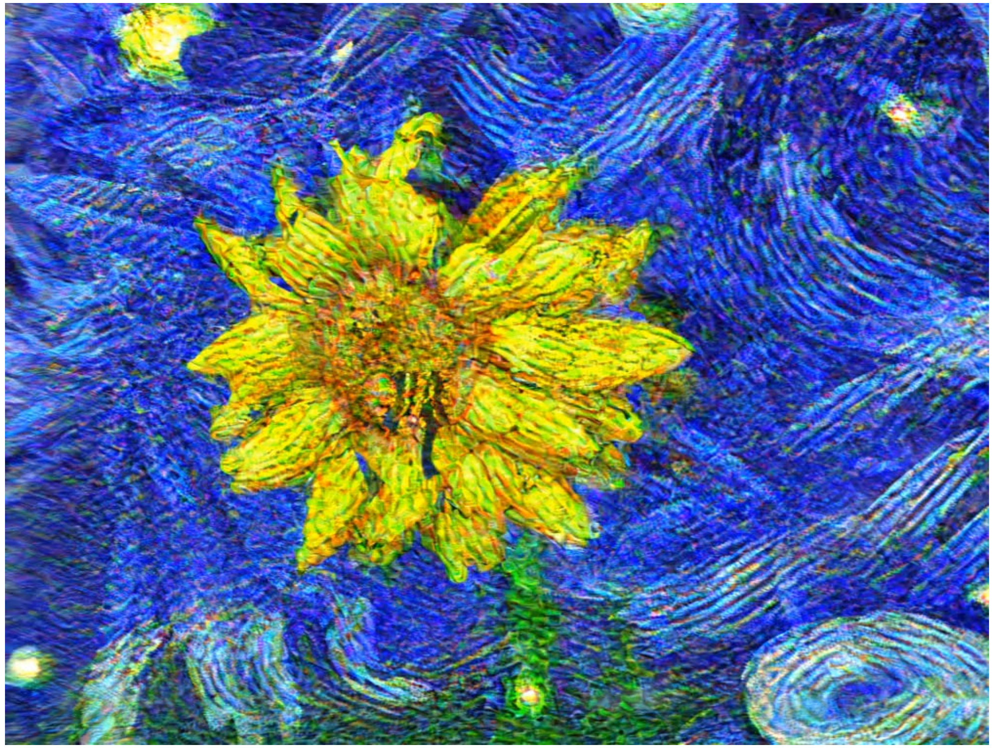
- The Hong Kong University of Science and Technology (Guangzhou)
- Ph.D. student in AI, advised by Prof. Ying-Cong Chen. 2022 - present

- Dalian University of Technology
- B.Sc. in Mathematics.   2018 – 2022

- I previously interned at NIO and MEGVII Technology.

NeRF: represent 3D scenes as neural nets

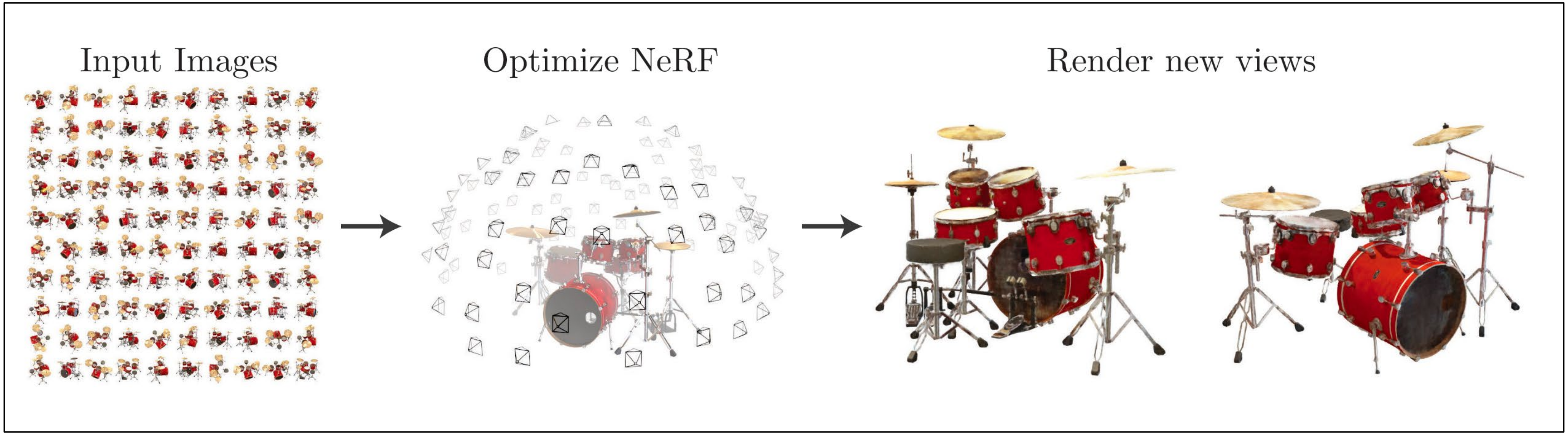
- NeRF: An implicit neural representation for 3D scenes.
- Application: novel view synthesis, reconstruction, generation, …



NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

NeRF: represent 3D scenes as neural nets

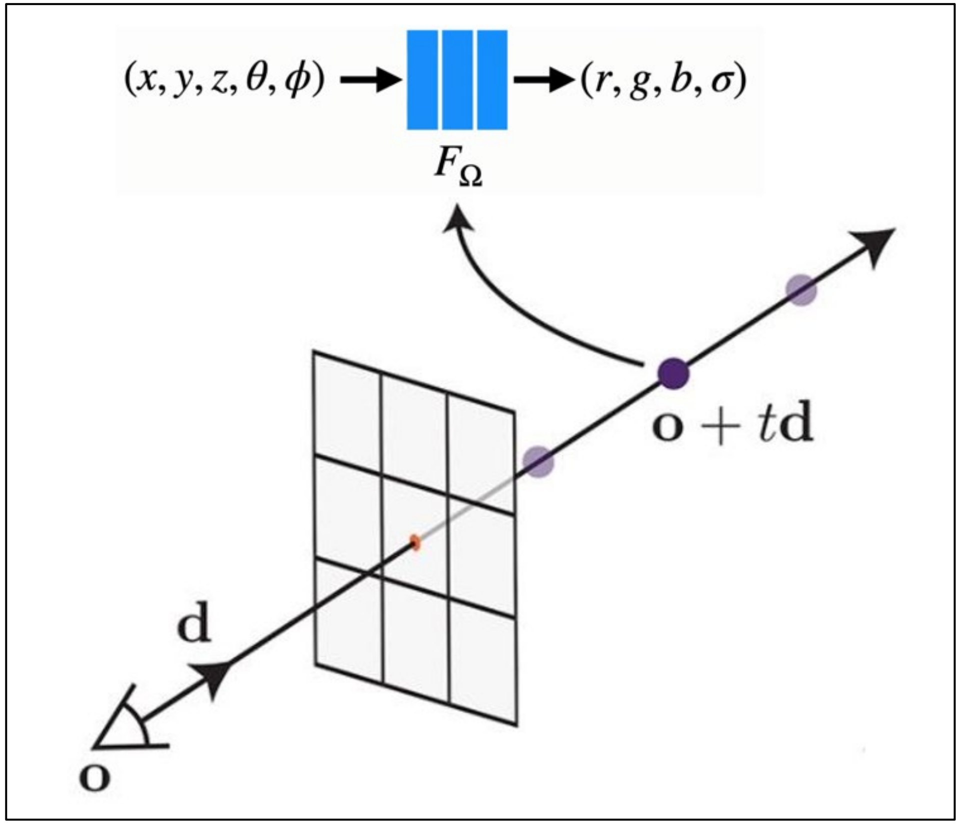
- Input: multi view images, intrinsic and extrinsic
- Training: optimize a MLP to fit the scene
- Inference: query the MLP to render novel view images
- Objective: Image similarity



NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

NeRF: represent 3D scenes as neural nets

- Ray casting: cast a ray from camera origin to pixel, then sample points from the ray.
- Volume rendering: mimic the 3D world as a “cloud”, each point in the “cloud” contribute its color.



Rendering model for ray r(t) = o + td:

$$C \approx \sum_{i=1}^{N} T_i \alpha_i c_i$$

weights

colors

How much light is blocked earlier along ray:

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_j)$$

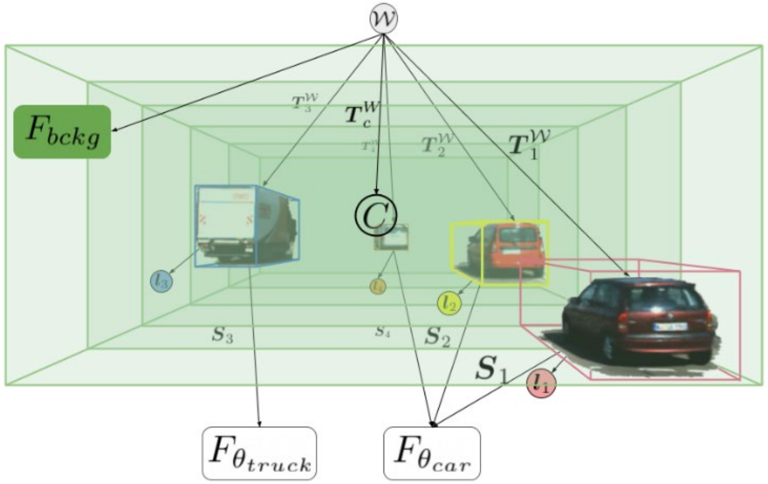
How much light is contributed by ray segment *i*:

$$\alpha_i = 1 - e^{-\sigma_i \delta t_i}$$

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

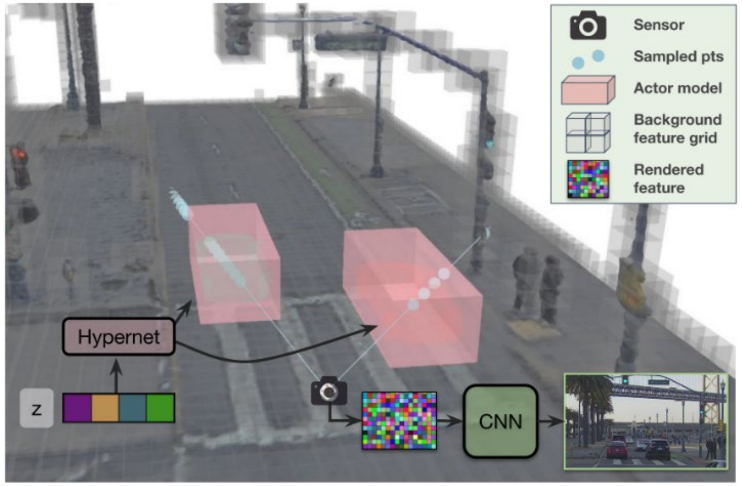
NeRF in AD: reconstruct the real world and replay it

- In recent years, the community has witnessed remarkable progress in NeRF-based driving scene simulation. These simulations display photorealistic reconstructions of our real world.



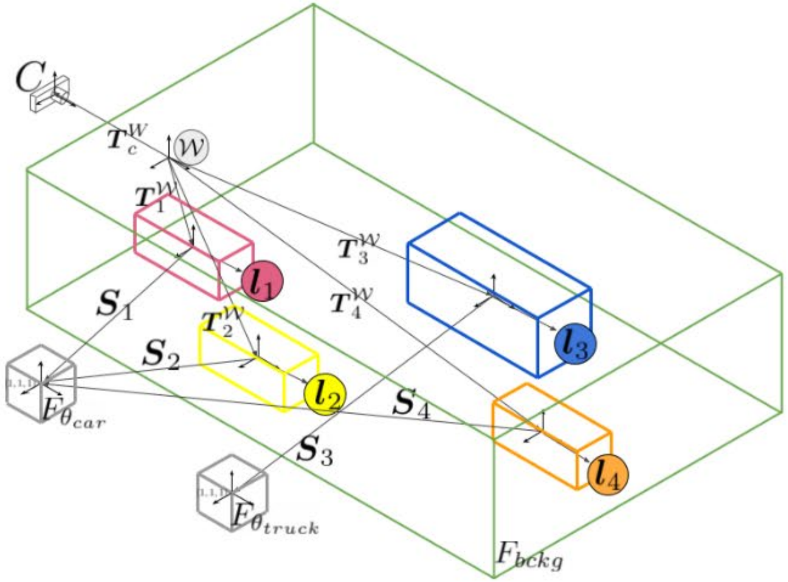
Neural Scene Graphs, CVPR 2021

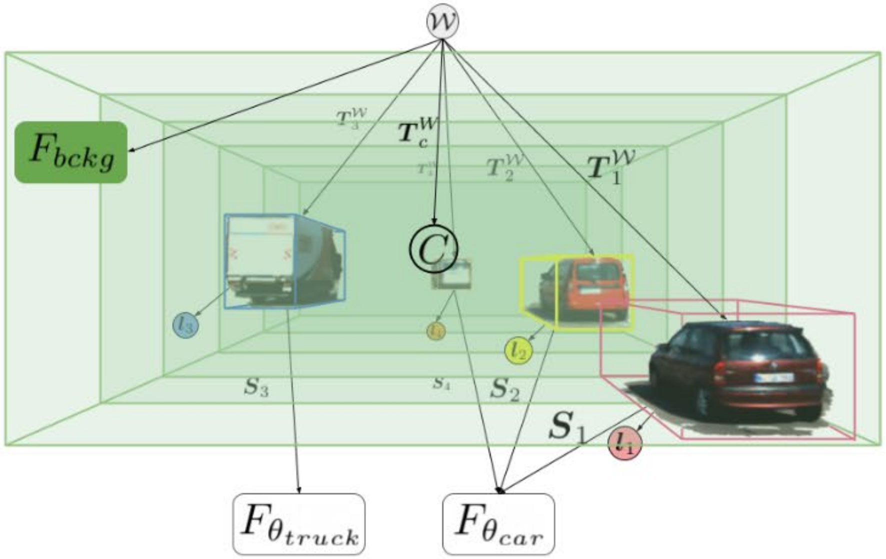
Block NeRF, CVPR 2022



UniSim, CVPR 2023

Neural Scene Graphs for Dynamic Scenes

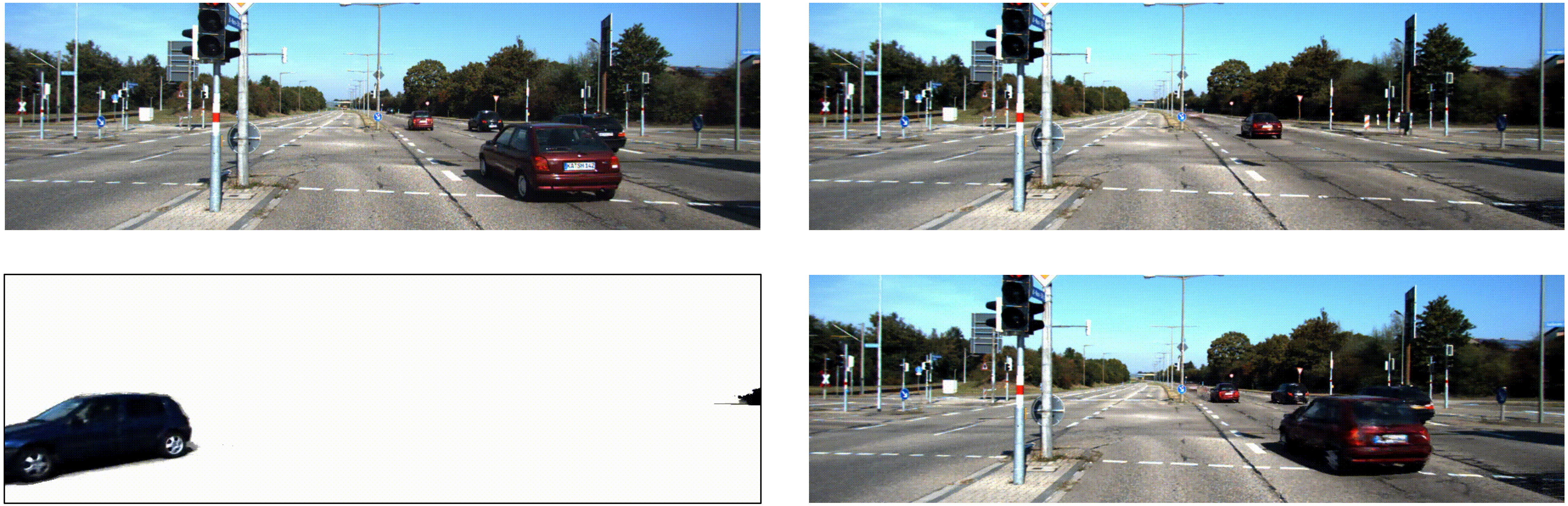




- The first exploration of NeRF in driving scenes.
- NSG disentangle dynamic objects and static background by explicit 3D boxes.
- The sequential 3D boxes are obtained from GT or detection+tracking

Neural Scene Graphs for Dynamic Scenes, CVPR 2021

Neural Scene Graphs for Dynamic Scenes



- NSG can control 6D pose of each object by changing the 3D box layout
- The 3D box layout is descripted by rotation and location of object in each frame

Neural Scene Graphs for Dynamic Scenes, CVPR 2021

Neural Scene Graphs for Dynamic Scenes

- NSG provides basic primitive (3D box) to decompose driving scenarios.
- Limitation:
- NSG generate 3D assets from pre-collected data. The scale of data is limited to the amount of real world captured data.
- What if we leverage generative model to synthesize unlimited data for free?

Neural Scene Graphs for Dynamic Scenes, CVPR 2021

Applications of Generative NeRF in autonomous driving

- Motivation:
- Generate free training data by AIGC (GAN, NeRF, diffusion…)
- Provide realistic evaluation and simulation

- Advantage:
- 1. No need for human annotation
- 2. Controllable (6D pose, lighting), easy to create long-tail scenes / corner cases
- 3. Nearly the same distribution with real world data, thus no need for domain adaptation
- 4. Photorealistic appearance compared with graphic engine (Unreal …)

Our work: use NeRF to synthesize training data

Lift3D: Synthesize 3D Training Data by Lifting 2D GAN to 3D Generative Radiance Field

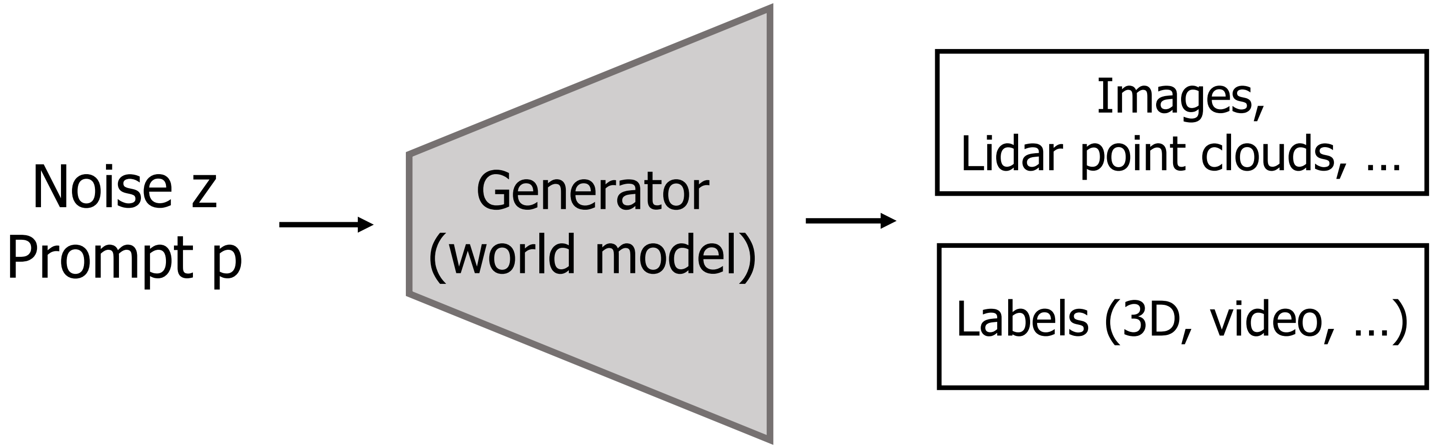
Leheng Li[1], Qing Lian[2], Luozhou Wang[1], Ningning Ma[3], Ying—Cong Chen[1,2]

[1]HKUST(GZ), [2]HKUST

[3]NIO

Project page: https://len-li.github.io/lift3d-web/

Imagine there is an AIGC algorithm that generate training data for free



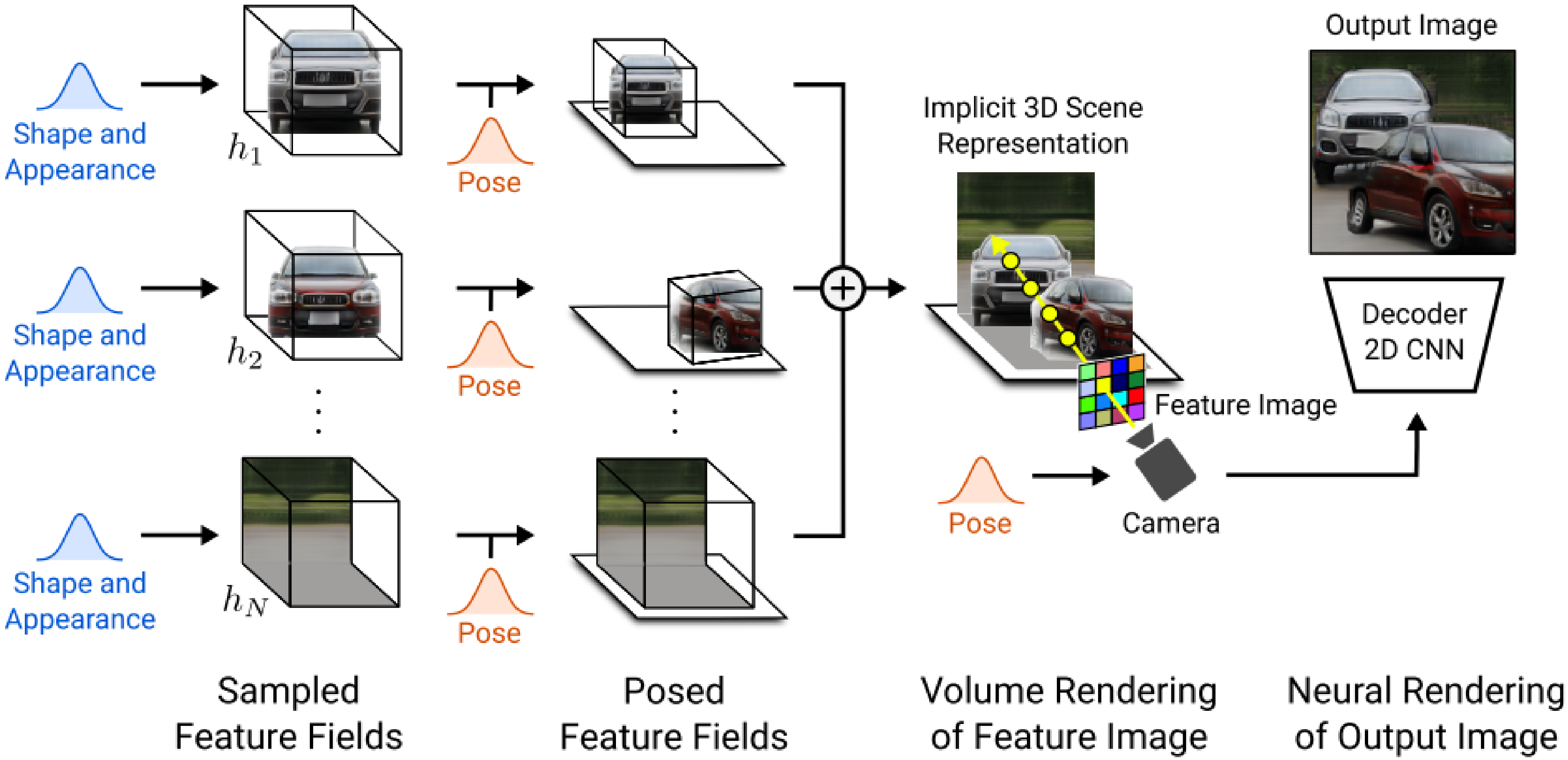
Evaluation setting: data augmentation

- It can be challenge to build a comprehensive model with world knowledge.
- Narrow the problem: synthesize objects and augment them to original scene.

Lift3D: Synthesize 3D Training Data by Lifting 2D GAN to 3D Generative Radiance Field, CVPR 2023

Baseline: GIRAFFE (CVPR 2021 best paper)

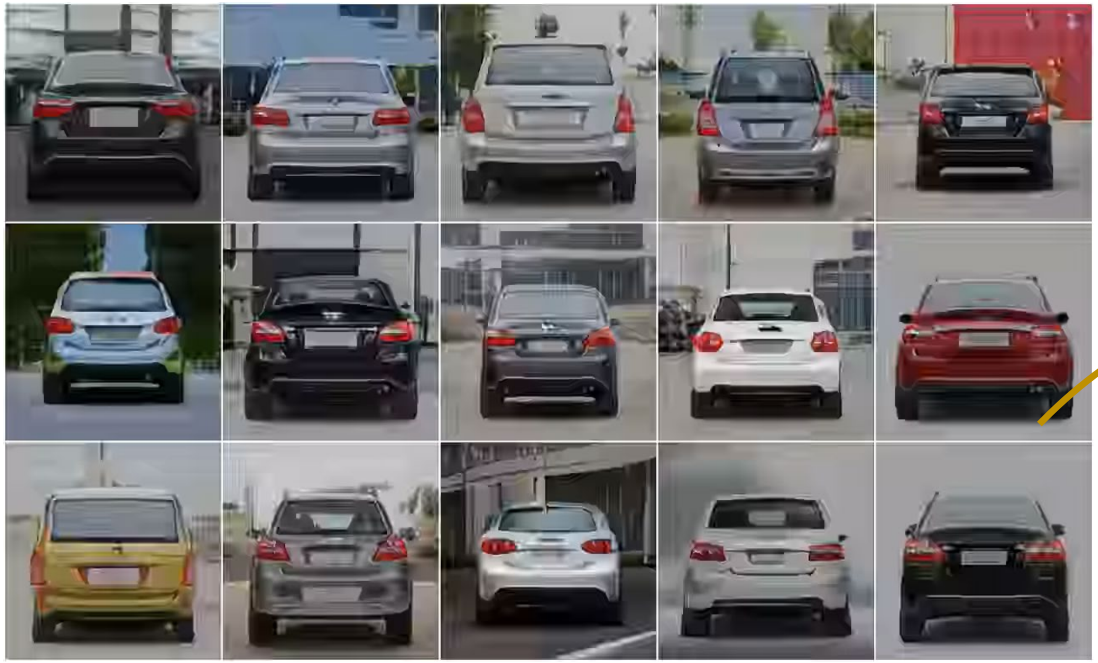
- Method: NeRF + GAN



GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, CVPR 2021

Use GIRAFFE to augment existing dataset

- Generate new objects and add them to existing scenes



Generated objects

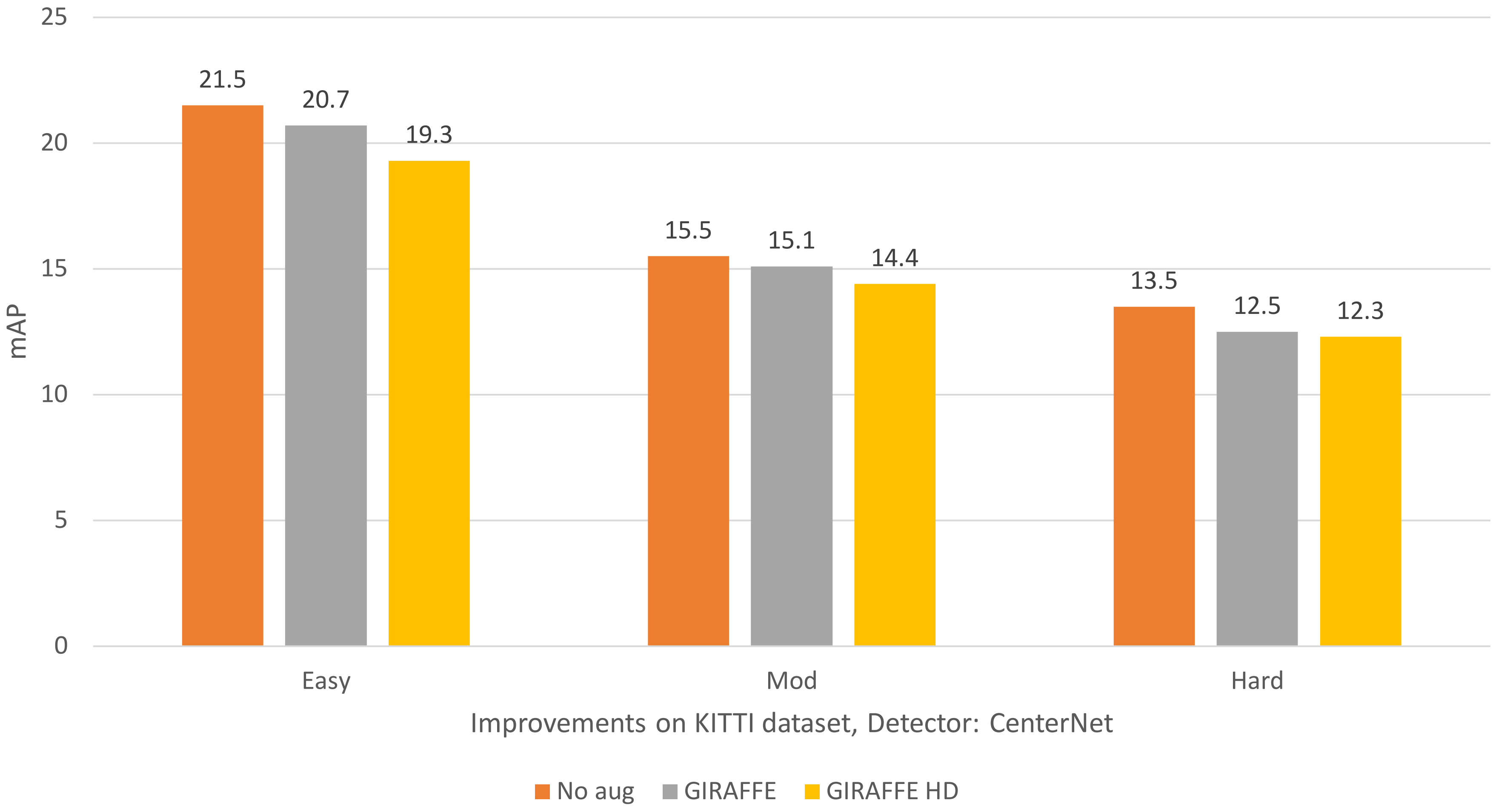
Add objects

nuScenes dataset

GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, CVPR 2021

Augmentation results of GIRAFFE

- Experiments: Impact of 3D detection accuracy on KITTI dataset



Augmentation results of GIRAFFE

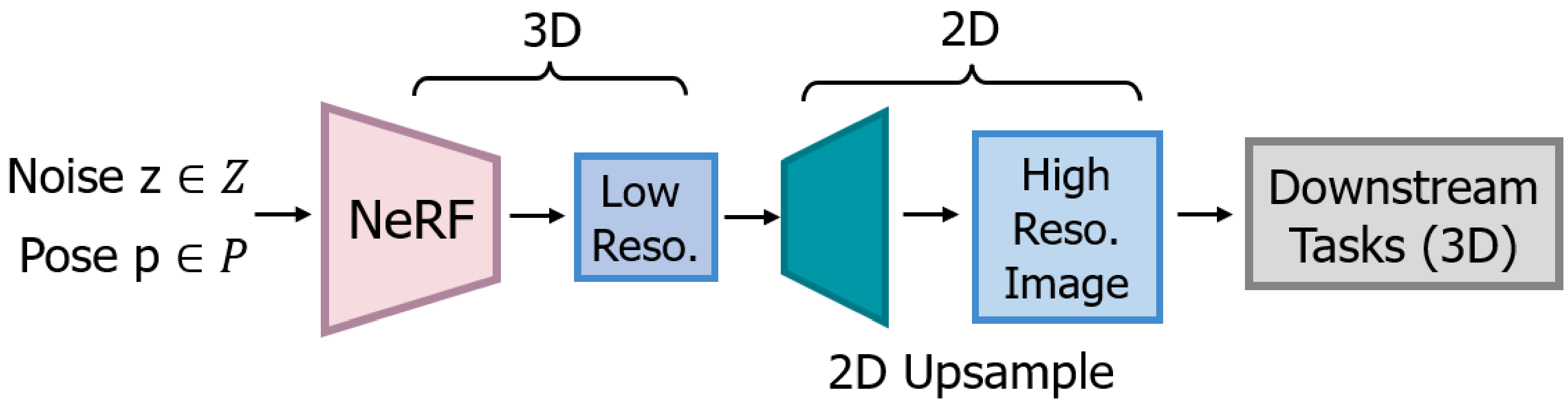
- Limitation: Augmentation of GIRAFFE introduce negative effect.
- Underlying mechanism: The generated images don't fit the given label

Generated multi-view images of an object by GIRAFFE

GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, CVPR 2021

Why previous method fall short of 3D consistent generation?

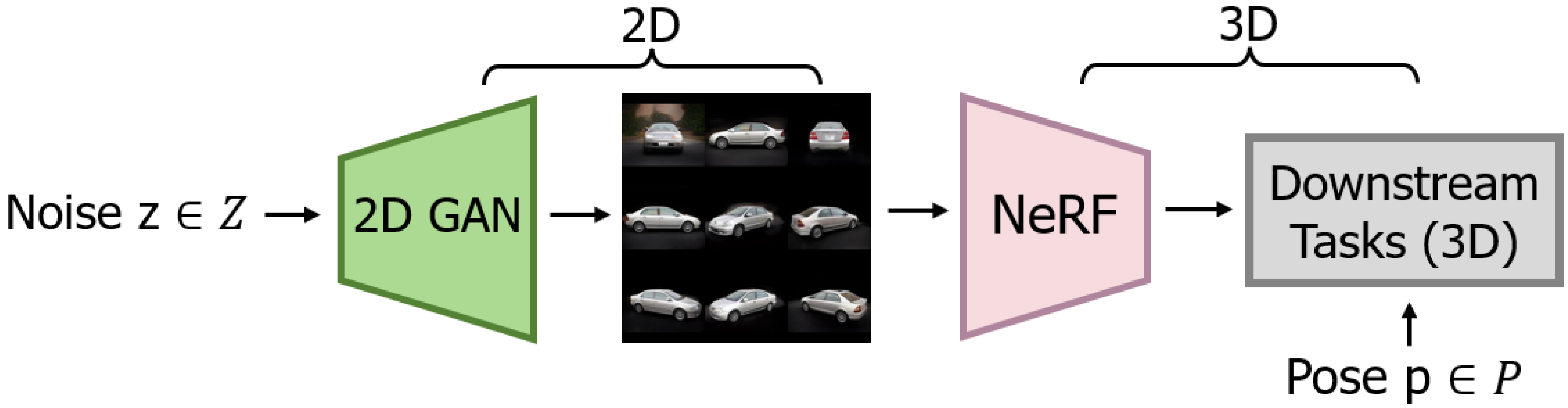
- Due to computational issue, generative NeRF typically adopt a two stage pipeline:
- 1. use volume render to generate the low resolution feature.
- 2. upsample the feature to the final image by 2D upsampler.
- Empirical results show that this pipeline does not strictly preserve 3D consistent synthesis due to 2D upsampler.



GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, CVPR 2021

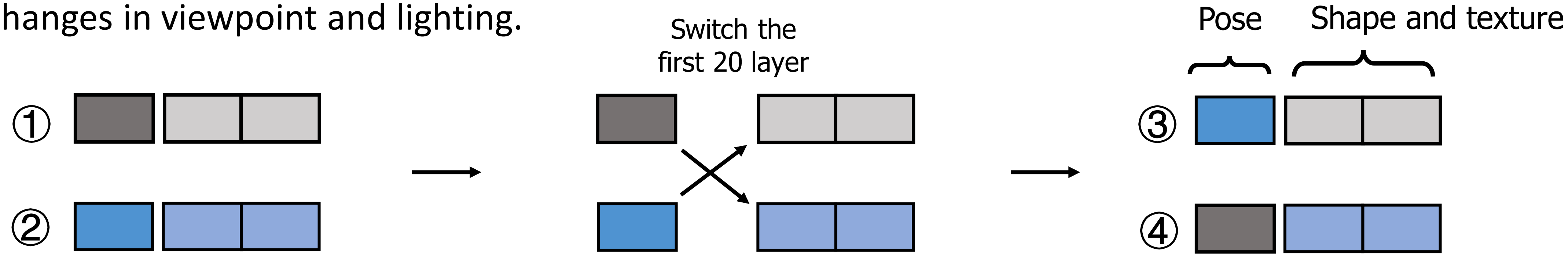
How to escape the computational bottleneck?

- Our method: Disentangle the 2D-3D generation.
- 2D GAN: image synthesis. NeRF: 3D synthesis
- Without relying on fixed-resolution 2D upsampler, Lift3D perform strict 3D consistent synthesis that generalize to any camera parameters.



Mechanism: GAN disentanglement

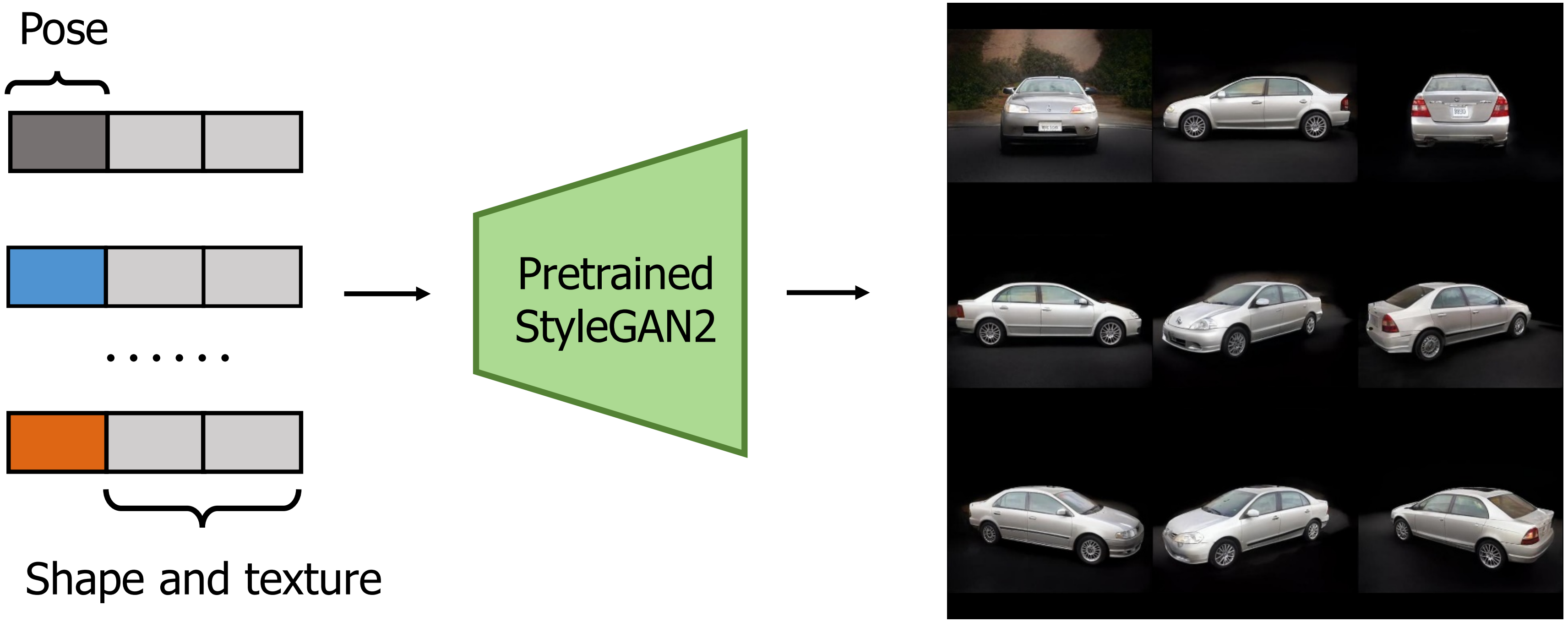
- Latent code: a high dimensional embedding that determine the content of image
- The latent space of GANs is found to be interpretable and controlled for image synthesis, allowing for changes in viewpoint and lighting.



GANSpace: Discovering Interpretable GAN Controls, NeurIPS 2020

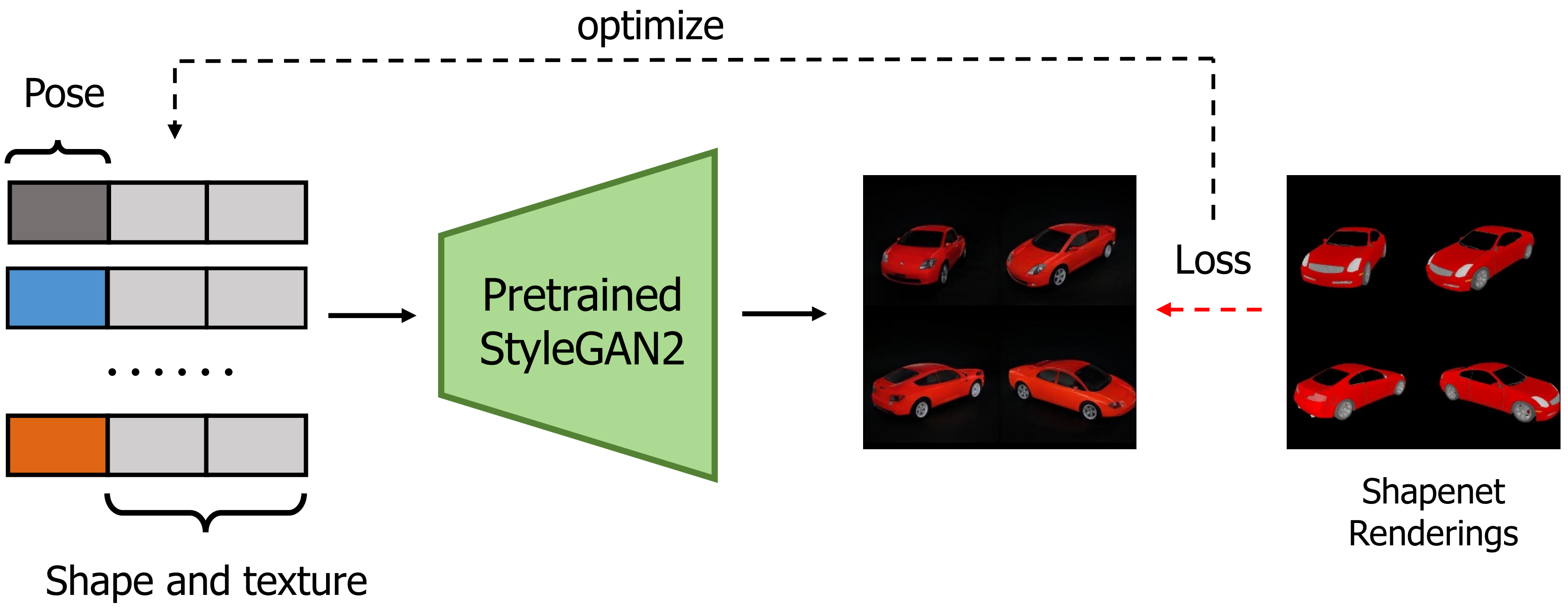
Two stage pipeline

- First stage: StyleGAN2 generates multi-view images of a specific object
- StyleGAN2 provides photorealistic synthesis + rough 3D controllability
- Disentangled 2D GANs allows to generate images with <u>3D pose label</u>



Two stage pipeline

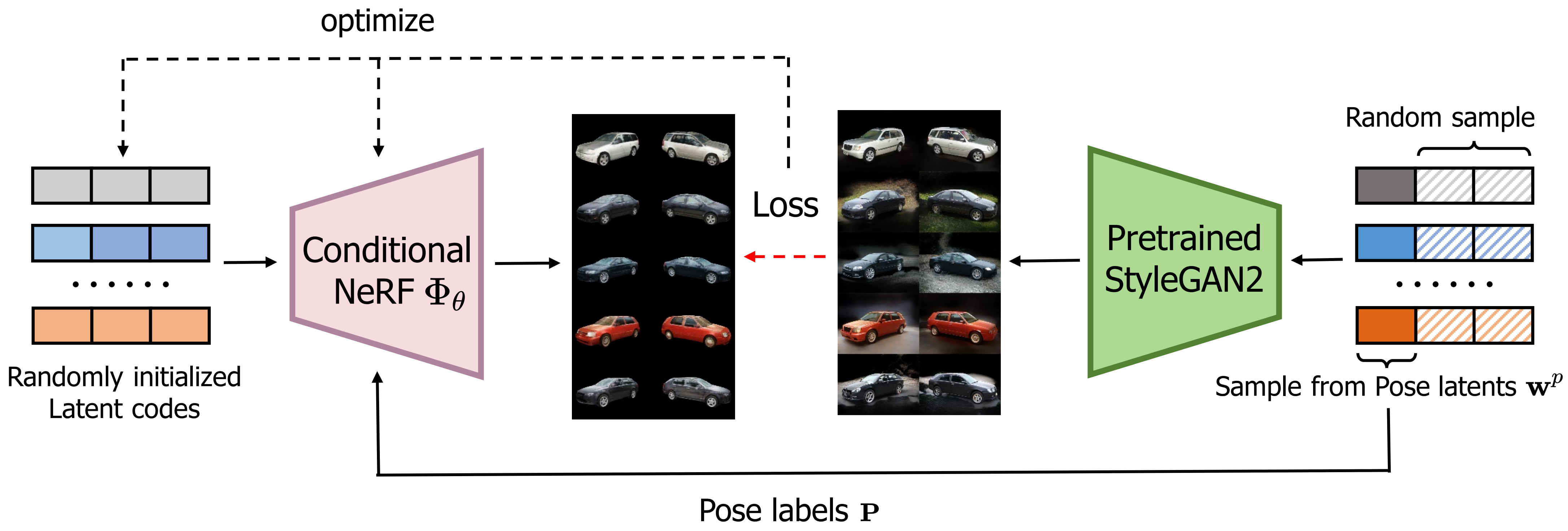
- First stage: StyleGAN2 generates multi-view images of a specific object
- Method: With the GT pose of synthetic data, we find pose latents by optimization



$$\hat{\mathbf{Z}}, \hat{\theta} = \arg\min_{\mathbf{z},\theta} \mathcal{L}(\mathbf{I}, \Phi_\theta(\mathbf{z}, \mathbf{P}))$$

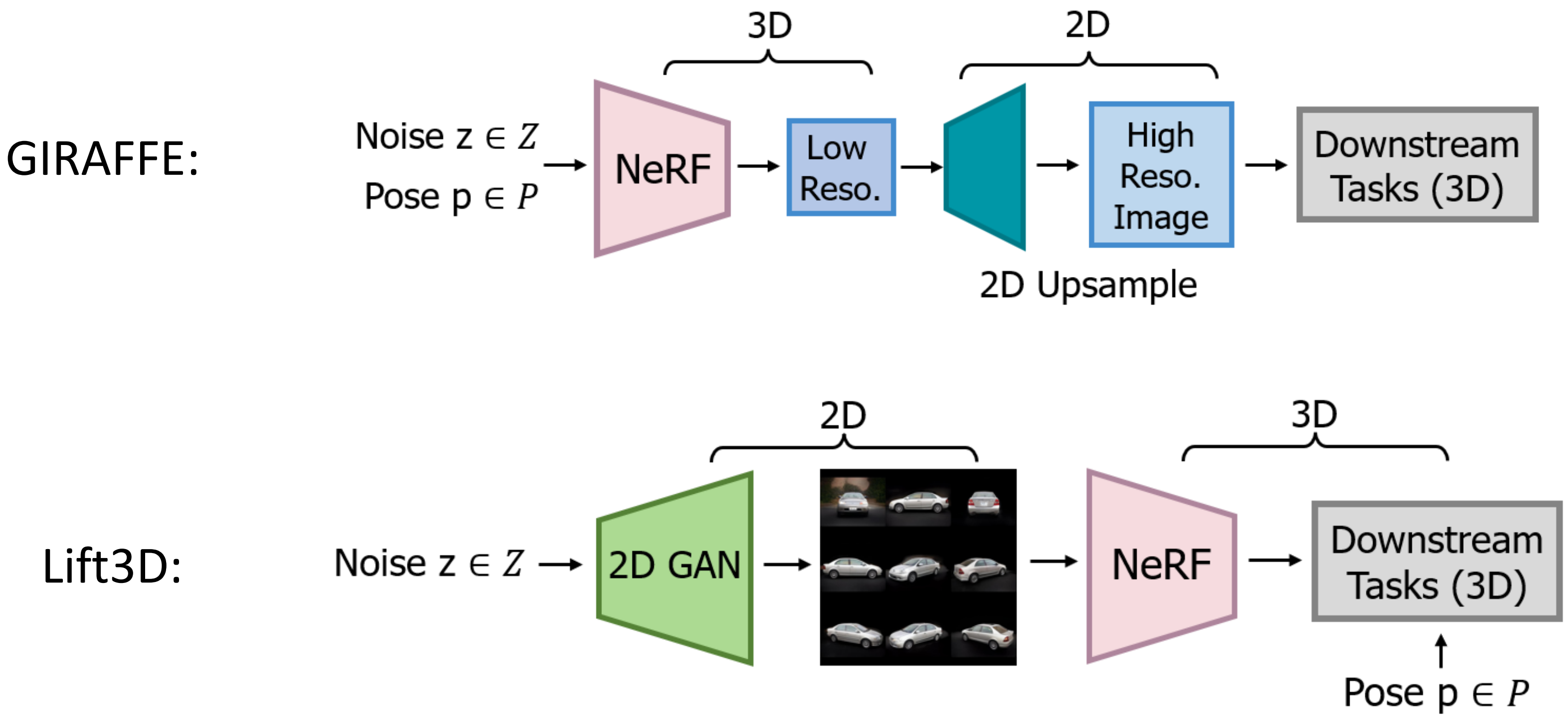
Two stage pipeline

- Second stage: Lift multi-view images to 3D NeRF.
- Conditional NeRF: All instances share the same NeRF network to encode prior.



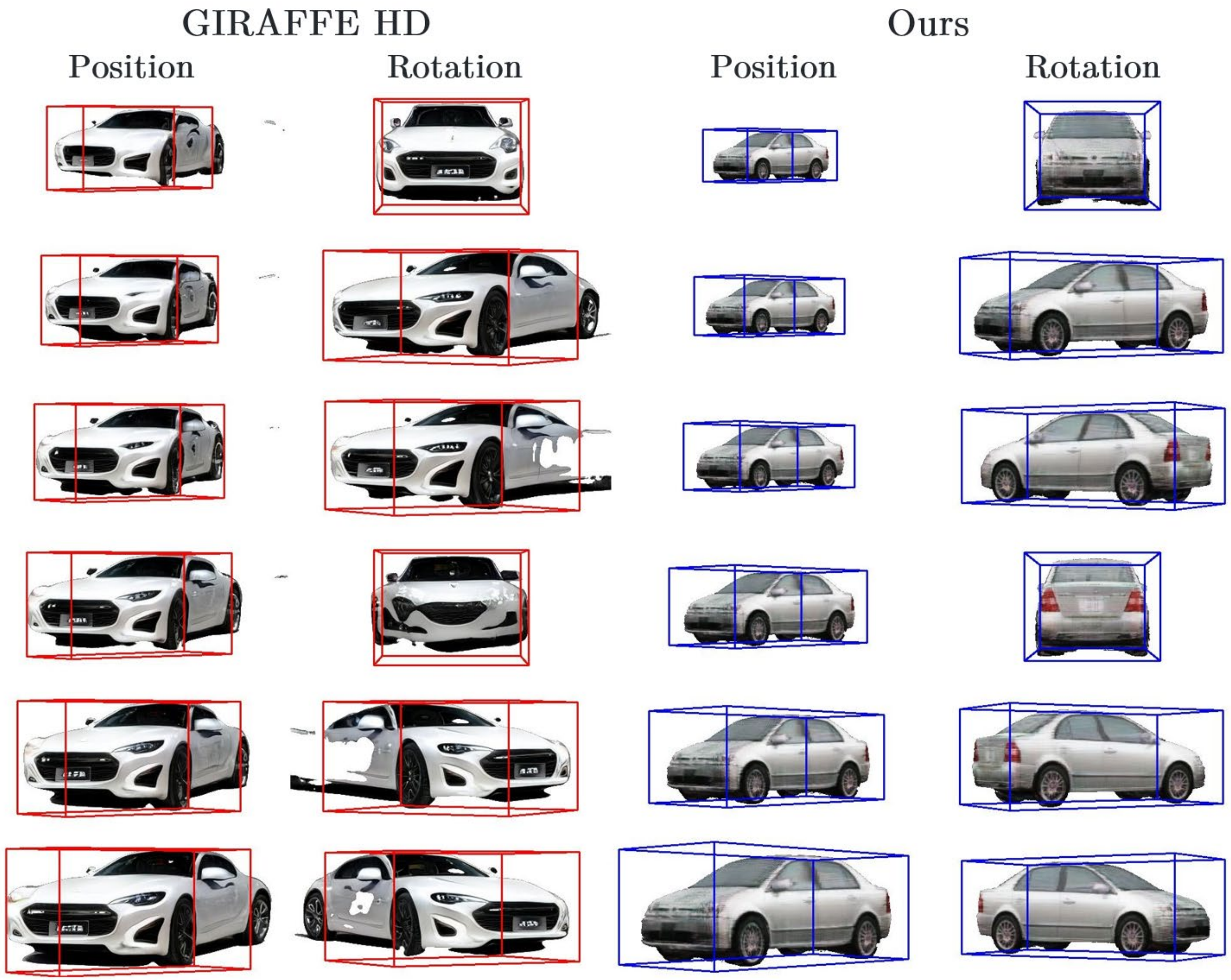
Mechanism

- GIRAFFE: 2D upsampler generalize poor to unseen pose
- Lift3D: disentangles 3D generation from image synthesis
- Our drawback: imperfect GAN disentanglement, NeRF reconstruction error, …



Results

- Visualization of multi-view synthesis with plotted 3D box



Composition

- Special design: The interaction of objects and environments.
- Shadow: casted from rounded rectangle,
- Map condition: objects are filtered by segmentation mask.


Input Image with Mask Prediction

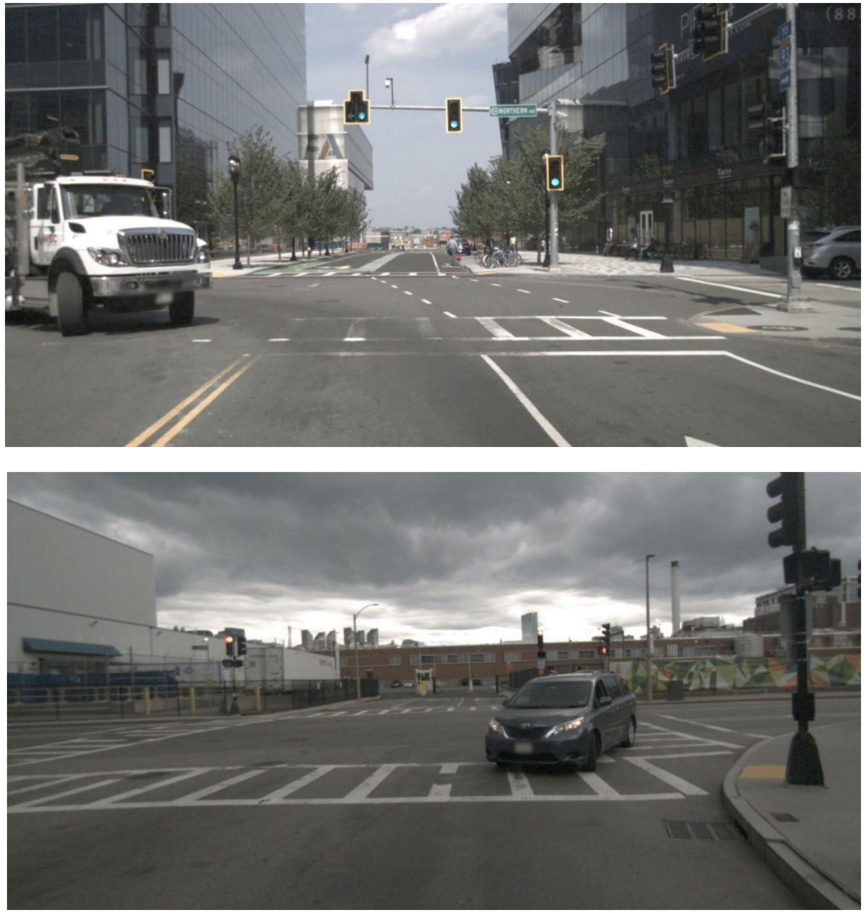

Augmented Image w/o Shadow, w/o Map


Augmented Image w/ Shadow, w/o Map


Augmented Image w/ Shadow, w/ Map

Results

- Visualization result of augmentation

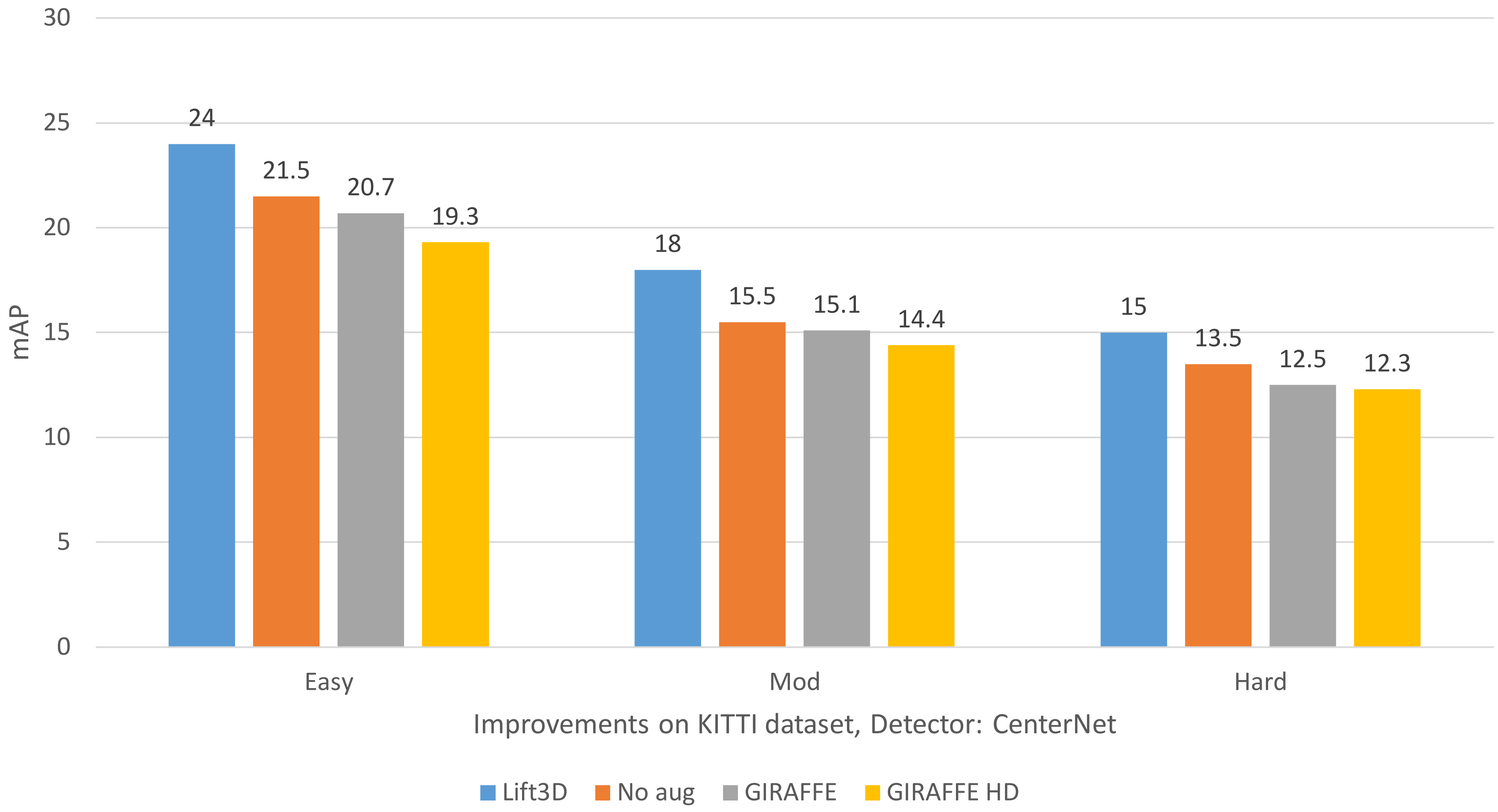


Original Dataset

Augmented Dataset

Results

- Improvement of 3D detection accuracy on KITTI dataset:



Summary

- Disentangled 3D generation provides tight 3D annotation
- Lift3D can synthesize images in any resolution by accumulating single-ray evaluation
- Without any domain adaptation, the generated data improves downstream task performance

Future work of AIGC in AD

- Generate long tail scenarios to enhance robustness
- Leverage generative prior to reconstruct real-world objects
- Trajectory generation: synthesize traffic flow
- Scene generation: closed-loop evaluation of self-driving car

Thanks for listening!
Q & A