LidaRF: Delving into Lidar for Neural Radiance Field on Street Scenes






Shanlin Sun[1] *Bingbing Zhuang*[3] *Ziyu Jiang*[3] *Buyu Liu*[3] *Xiaohui Xie*[1] *Manmohan Chandraker*[2,3]

[1] UCI University of California, Irvine

[2] UC San Diego

[3] NEC NEC Laboratories America

Problem Statement

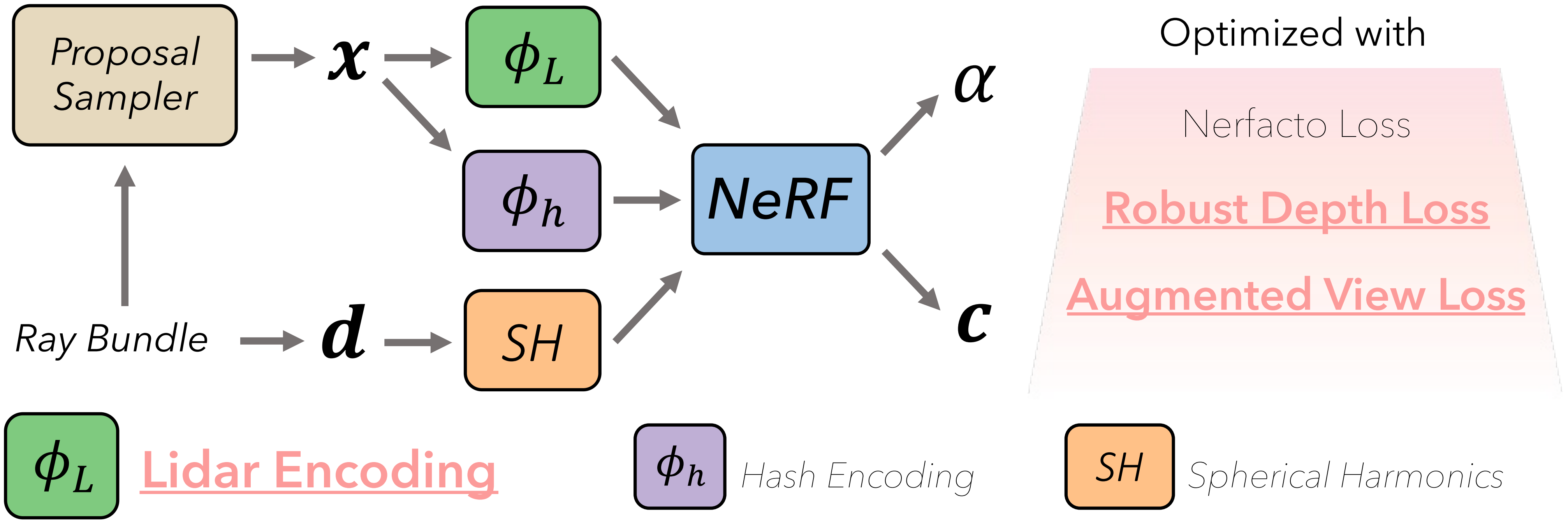
Input:

From RGB camera, Lidar, sensor poses

Goal:

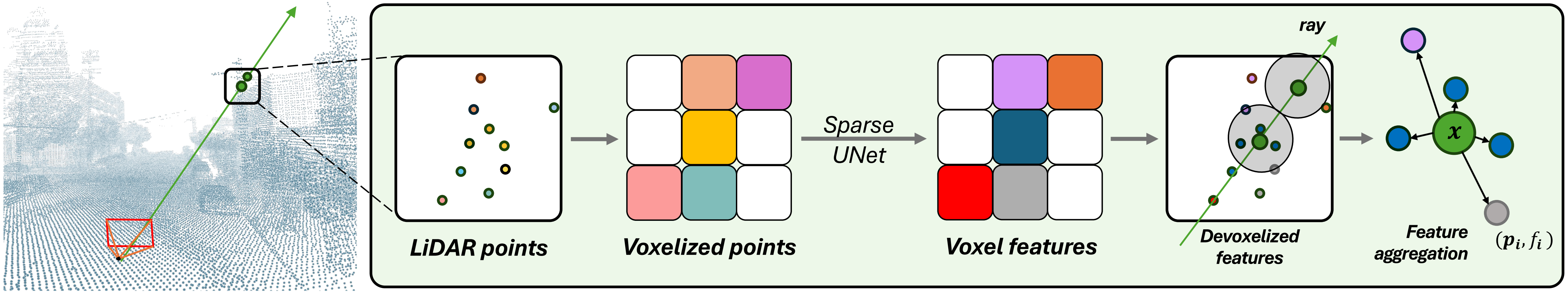
Photorealistic appearance simulation of street scenes for training and verification of autonomy

Our Contributions



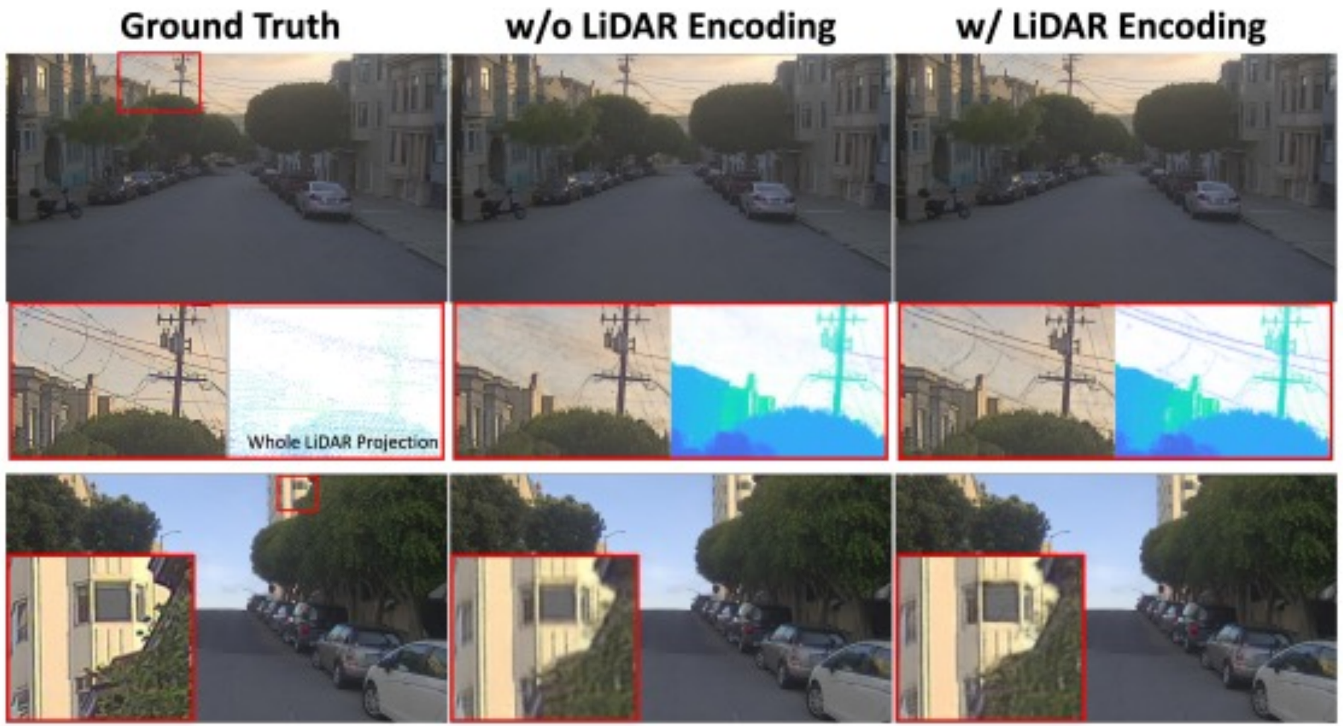
Motivation: Modern autonomous systems are often equipped with Lidar.
How can we use it more than just as a depth loss?

Contribution #1: Lidar Encoding



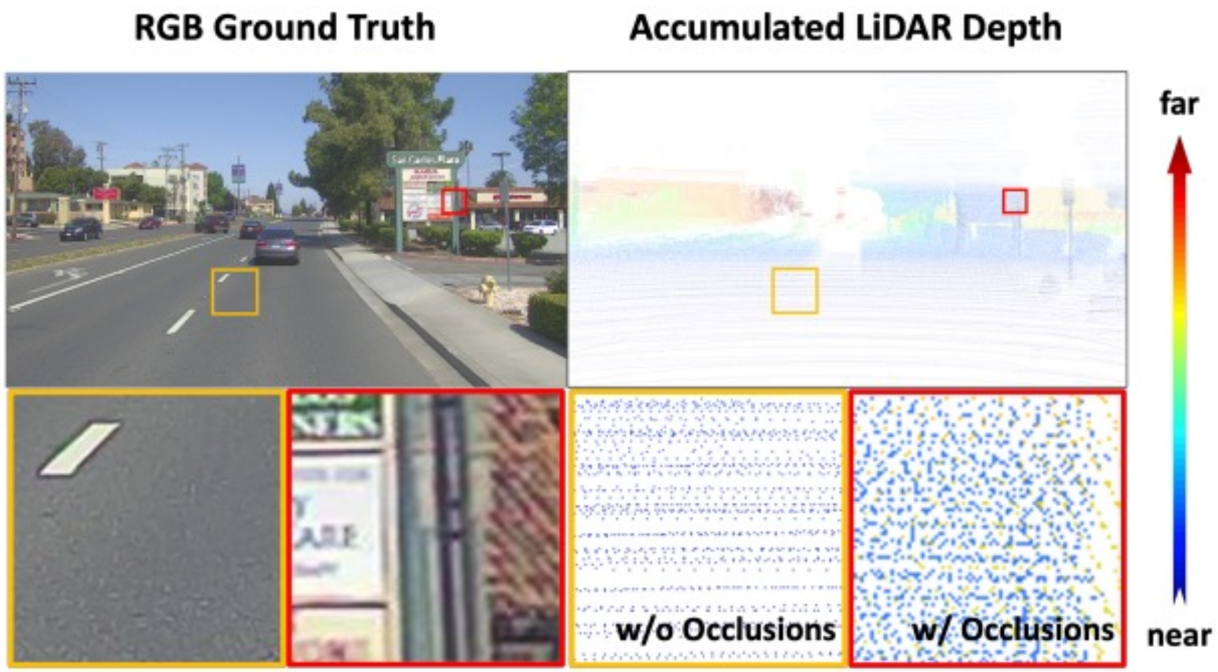
- Lidar holds strong potential for geometric guidance
- Lidar encoding through 3D sparse CNN has proven powerful in 3D perception framework, but is underexplored in NeRF
- Fuse Lidar encoding and hash grid feature

Lidar Encoding
Ablation Study



Methods	Interpolation			Lane Shift	
	PSNR↑	SSIM↑	LPIPS↓	FID↓ @ 2m	FID↓ @ 3.7m
Original Hash	27.090	0.804	0.247	110.0	131.7
Double Hash	27.153	0.808	0.234	109.3	132.1
MLP	27.119	0.805	0.246	108.0	131.6
PointNet++	27.076	0.804	0.247	108.7	131.2
Ours	**27.219**	**0.810**	**0.228**	**105.6**	**128.7**

Contribution #2: Robust Depth Supervision



Issue:

Inter-points occlusion due to the camera-Lidar displacement

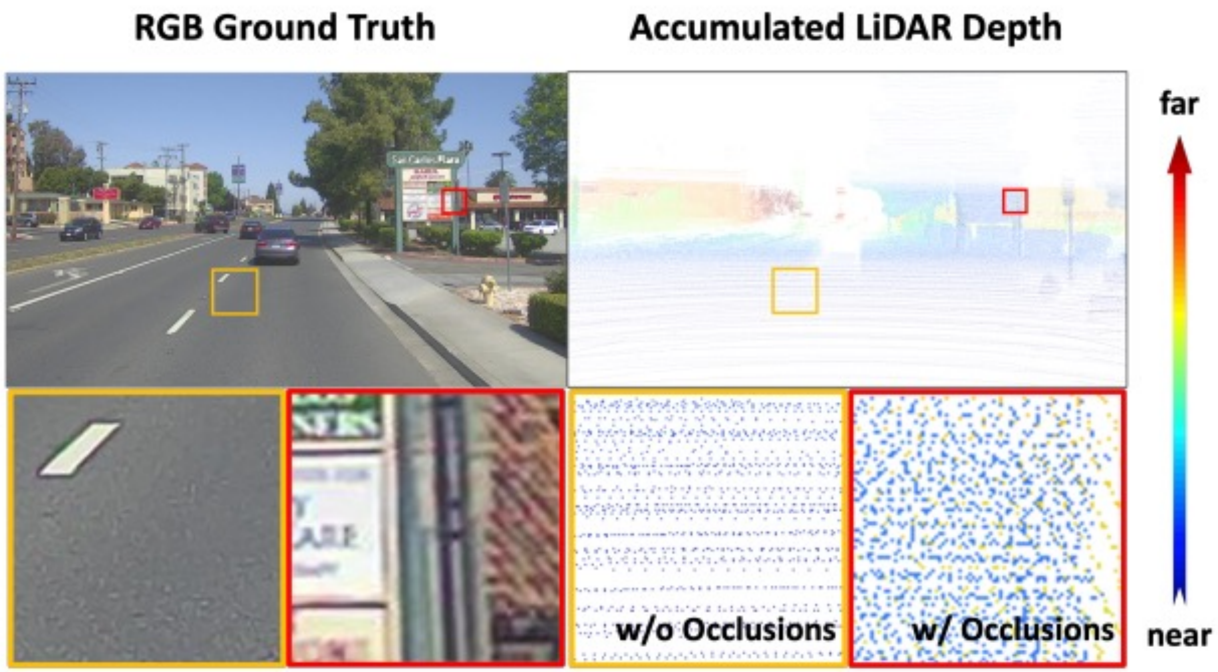
Goal:

Discard fake depth supervision adaptively

Intuition:

Count on near points initially and gradually add far points during training

Contribution #2: Robust Depth Supervision



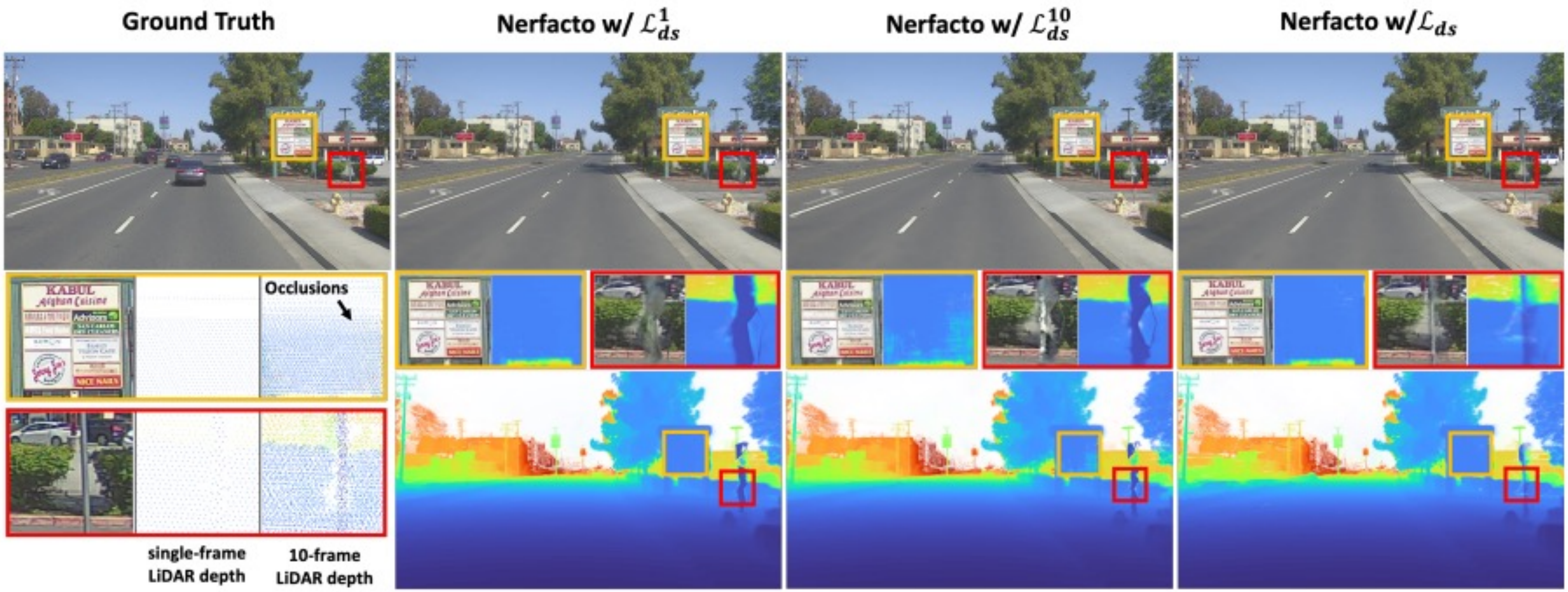
$$\mathcal{D}^m_{\text{reliable}} = \{\mathcal{D}_i \mid \mathcal{D}_i \leq \epsilon^m_t,\ \ \mathcal{D}_i \leq \hat{\mathcal{D}}_i + \epsilon^m_o,\ \ \mathcal{D}_i \in \mathcal{D}\},$$
$$\epsilon^m_t = \min\{\alpha_t \epsilon^{m-1}_t,\ \epsilon_t\}, \quad \alpha_t > 1,$$
$$\epsilon^m_o = \max\{\alpha_o \epsilon^{m-1}_o,\ \epsilon_o\}, \quad \alpha_o < 1.$$

Curriculum learning:

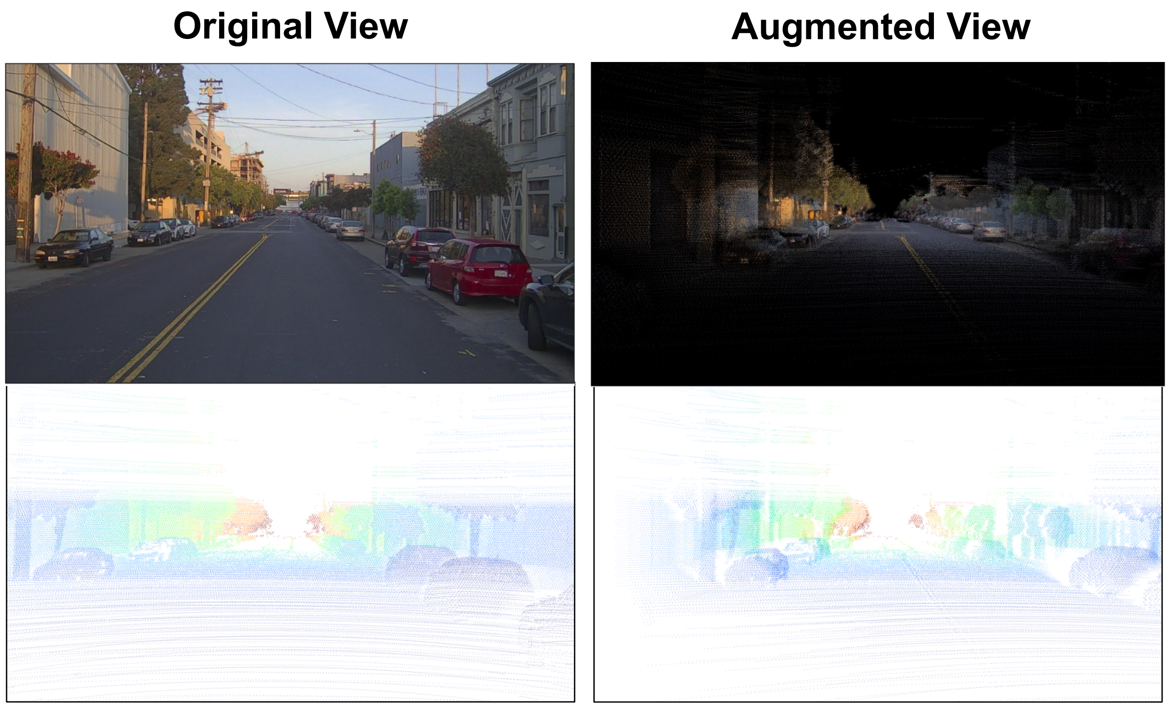
- Valid depth threshold ϵ^m_t increases at a rate of α_t
- Valid depth offset ϵ^m_o decreases at a rate of α_o
- Adopt URF loss for samples in $\mathcal{D}^m_{\text{reliable}}$

Robust Depth Supervision

Ablation Study



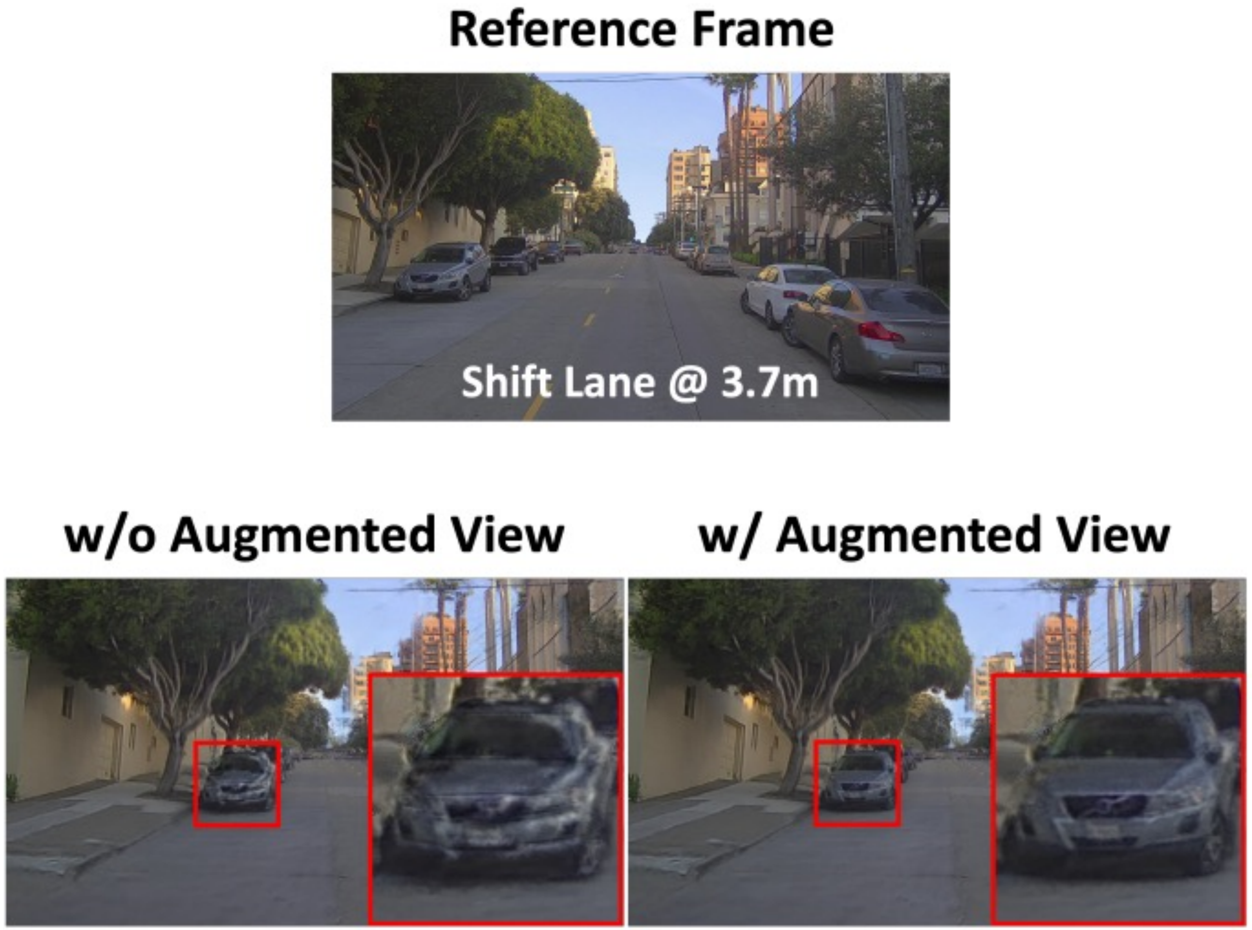
Contribution #3: **Augmented View Supervision**



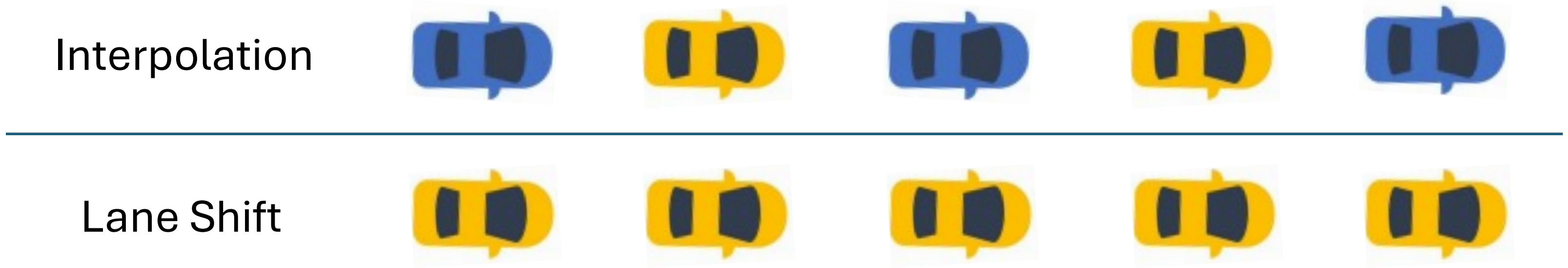
1. Colorize Lidar points in each Lidar frame
2. Accumulate colorized Lidar points
3. Project them to the augmented views
4. Filter out occluded signal (RGB / Depth)

Augmented View Supervision

Ablation Study



Compare with SoTA
PandaSet Dadaset

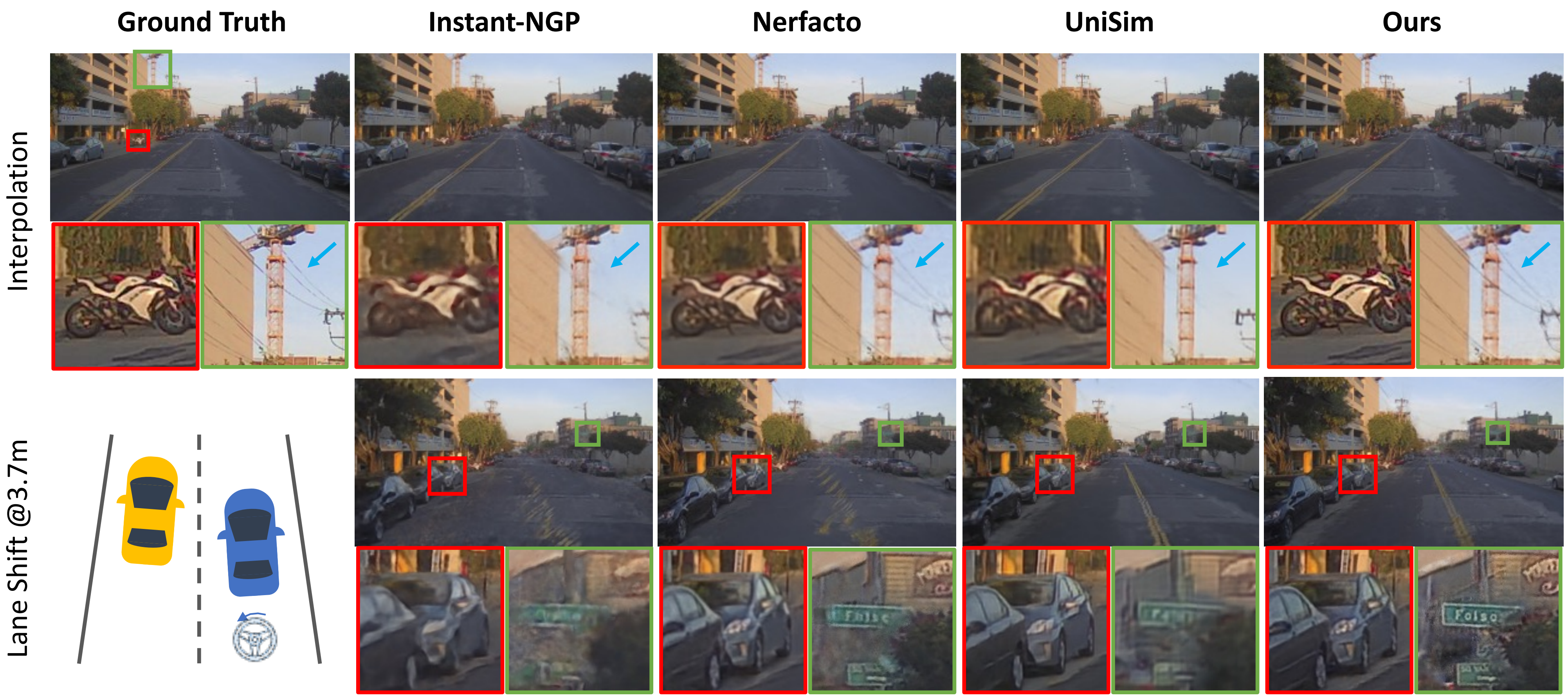


Blue: Training Views

Yellow: Test View

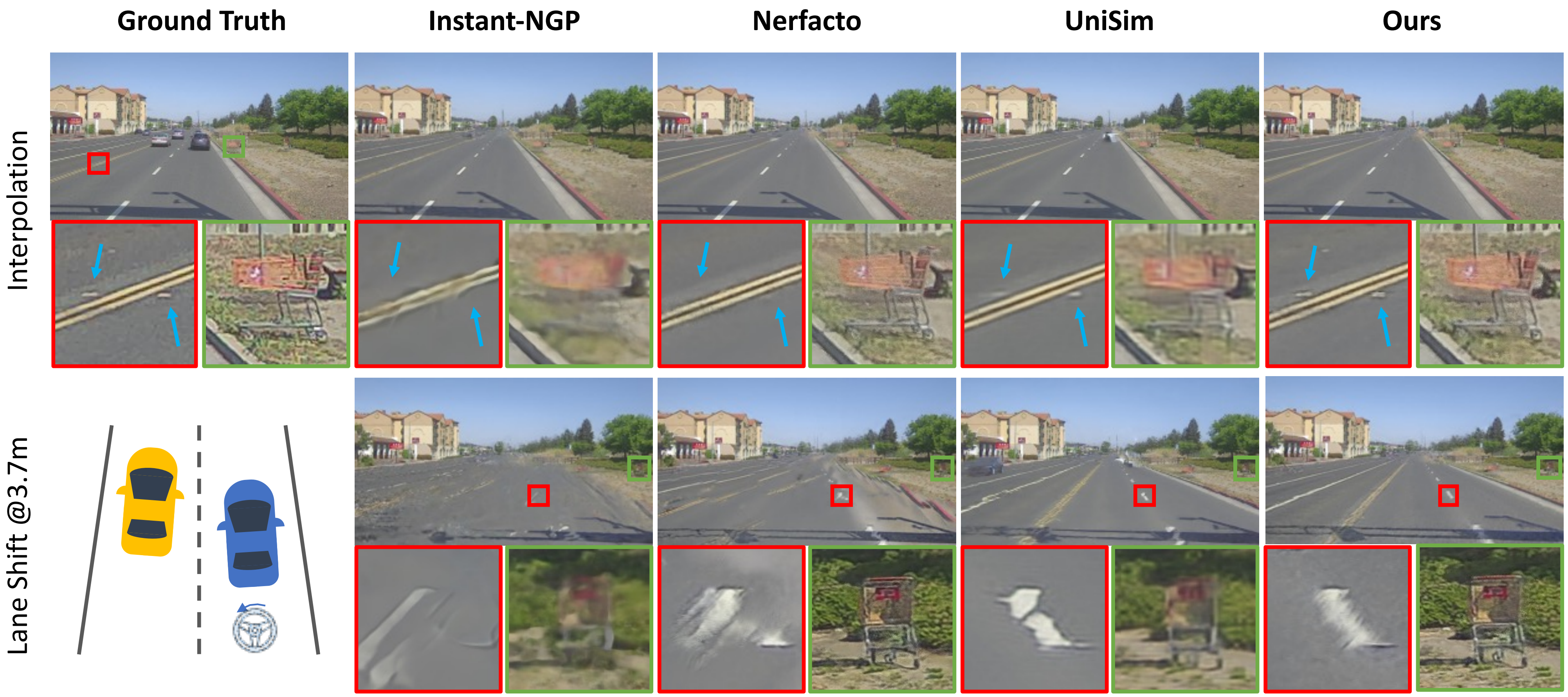
Methods	Interpolation			Lane Shift	
	PSNR↑	SSIM↑	LPIPS↓	FID↓ @ 2m	FID↓ @ 3.7m
Instant-NGP	24.282	0.733	0.408	140.3	173.2
Mip-NeRF 360	23.693	0.691	0.496	189.4	231.1
Nerfacto	27.122	0.804	0.268	116.7	151.0
UniSim	26.014	0.768	0.342	118.5	141.3
Ours	27.255	**0.812**	**0.224**	**106.5**	**126.0**

Compare with SoTA
PandaSet Dadaset



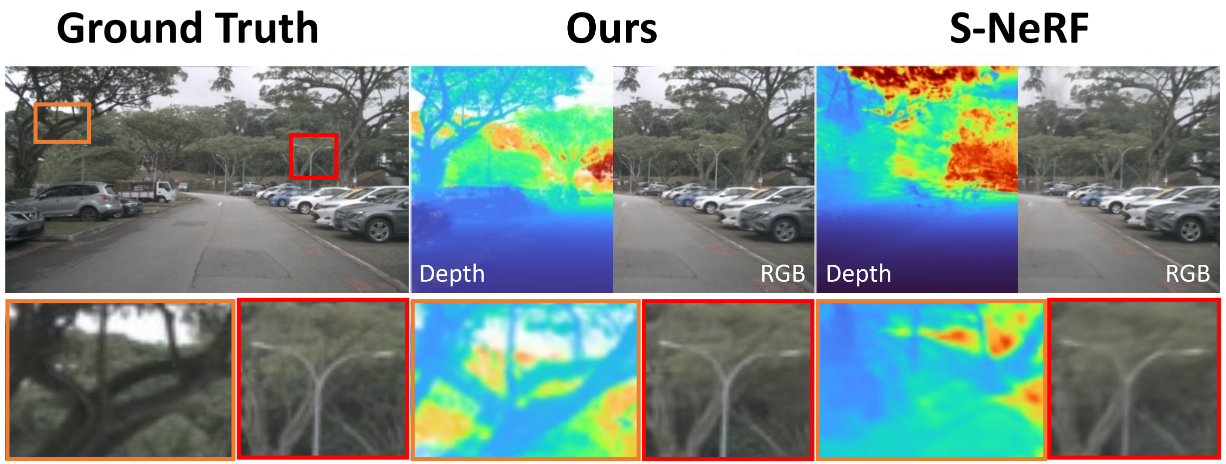
Compare with SoTA

PandaSet Dadaset



Compare with SoTA

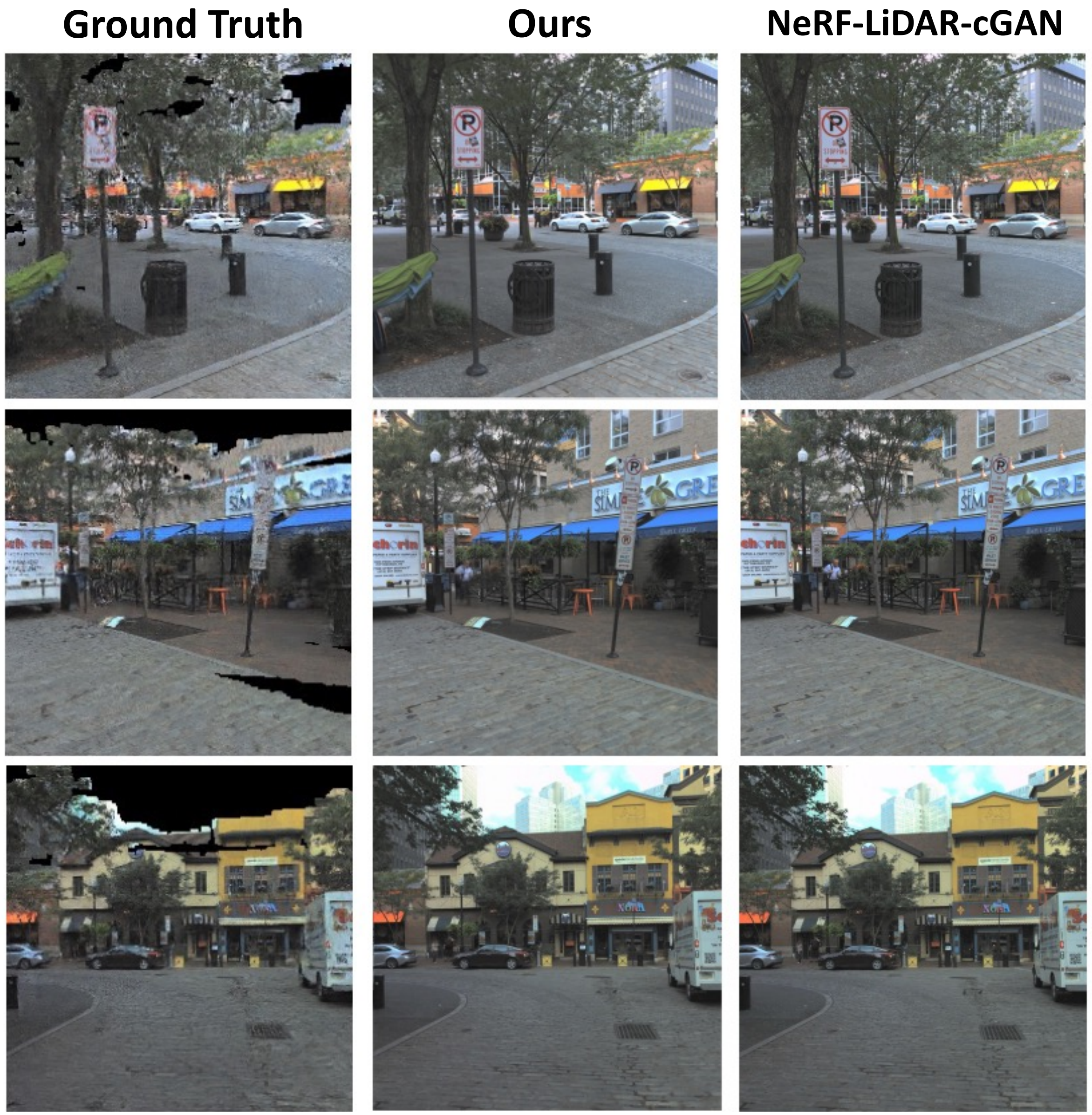
NuScenes Dadaset



Methods	S-NeRF	Ours			
		w/o $\mathcal{L}_{\text{ds}}$	w/o ϕ_L	w/o $\mathcal{L}_{\text{aug}}$	Full
PSNR↑	29.377	30.629	31.001	31.133	**31.162**
SSIM↑	0.859	0.871	0.873	0.883	**0.884**
LPIPS↓	0.349	0.278	0.237	0.222	**0.211**

Compare with SoTA

Argoverse Dadaset



Limitations & Future Work

- Not model dynamic objects in the scene
  → NeRF composition
- Not a real-time simulator
  → Textured Mesh / Gaussian Splatting
- Not good for totally unseen regions
  → Diffusion priors

Thanks!