Reliability-Aware Monocular Depth Supervision for Sparse-View Neural Reconstruction

Wayne Chu
waynechu@stanford.edu

Yashasvini Gopalan
ygopalan@stanford.edu

Changju Yuan
ycj2003@stanford.edu

Introduction

- Novel view synthesis reconstructs 3D scenes from posed images and renders unseen viewpoints.
- It is important for autonomous driving simulation, robotics, AR, and digital twins.
- NeRF and 3D Gaussian Splatting achieve strong rendering quality with multi-view images.
- Outdoor driving scenes are especially challenging because cameras usually move along a narrow forward-facing trajectory.

Problem Statement

- Sparse-view reconstruction is under-constrained when trained with RGB supervision alone.
- Limited overlap and weak parallax can lead to inaccurate geometry, floaters, and unstable depth.
- Monocular depth estimators provide dense geometric priors, but their predictions are noisy and scale-ambiguous.
- Applying depth supervision everywhere may hurt reconstruction if the depth prior is unreliable.
- We study whether reliability-aware monocular depth supervision can improve sparse-view NeRF and 3DGS reconstruction.

Dataset

KITTI (Seq. 02): Outdoor driving scenes with LiDAR depth and calibrated poses

Mip-NeRF 360 (Bicycle): Sparse outdoor object-centric scene for novel view synthesis

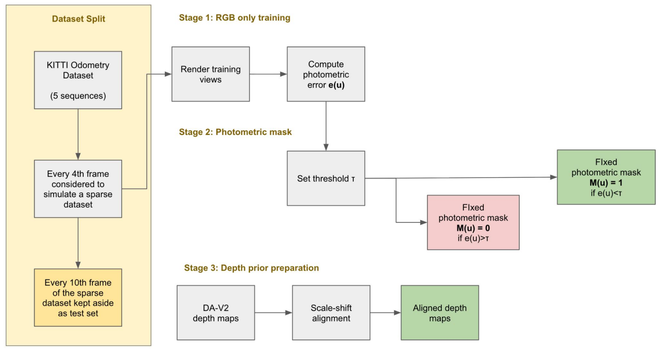

Fig 1: Dataset and depth prior preparation

Methods

Depth Alignment

Align DA-V2 relative depth to metric scale using LiDAR depth (KITTI) or sparse COLMAP points (Bicycle)

Photometric Masking

Train an RGB-only baseline and compute photometric error:

$$e(u) = \frac{1}{3}\sum |\hat{I}_c(u) - I_c(u)|$$

$$L_{\text{rgb}} = \|\hat{I} - I\|_2^2$$

$\hat{I}, I$: rendered and ground-truth RGB images

Generate a fixed binary mask using threshold τ:

$$M(u) = \begin{cases} 1, & e(u) < \tau \\ 0, & \text{otherwise} \end{cases}$$

NeRF & 3DGS Training

Apply masked MSE loss between the rendered depth and aligned DA-V2 depth prior:

$$L_{\text{depth}} = M(u)\,(\hat{d} - d)^2$$

Optimize:

$$L = L_{\text{rgb}} + \lambda_{\text{depth}} L_{\text{depth}}$$

$\hat{d}, d$: rendered depth and aligned DA-V2 depth prior
$M(u)$: fixed photometric mask

Experiments & Analysis

Experiments

- KITTISeq02 sparse-view setting + Bicycle scene
- Backbones: Mip-NeRF-360 and Splatfacto
- Depth prior: scale-aligned Depth Anything V2
- Ablation: mask threshold τ and depth weight λ

Scene / Model	RGB-only	Best Depth Supervision	Takeaway
KITTI / Mip-NeRF-360	PSNR 20.378, RMSE 2.703	PSNR 20.607, RMSE 3.580	Small PSNR gain, worse geometry
KITTI / Splatfacto	PSNR 14.903, RMSE 0.542	PSNR 15.932, RMSE 0.100	Clear improvement
Bicycle / Splatfacto	PSNR 17.731, RMSE 1.479	PSNR 17.036, RMSE 0.722	Geometry improves, RGB drops

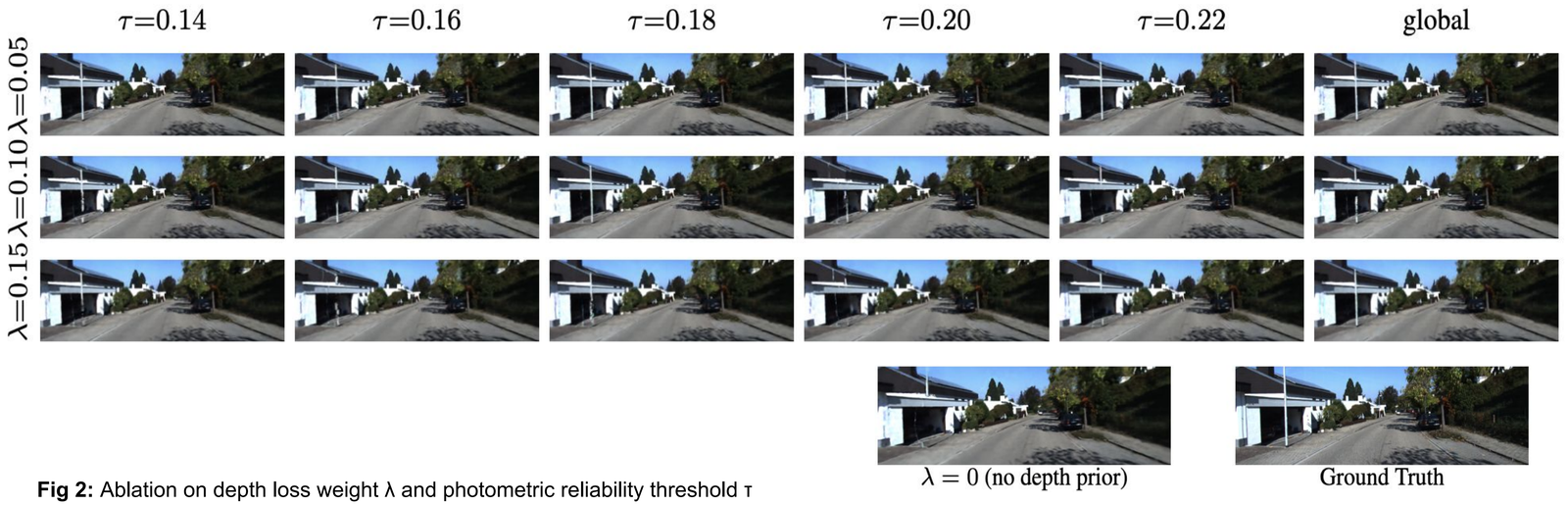

Fig 2: Ablation on depth loss weight λ and photometric reliability threshold τ

Conclusions & Future Work

Depth supervision improves geometry in some cases, but may hurt RGB rendering quality when over-weighted.

Future Work

- Better depth scale alignment
- Uncertainty-aware masks
- Adaptive depth-loss weighting
- Evaluation on more scenes and depth backbones