THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

ICCV 2025

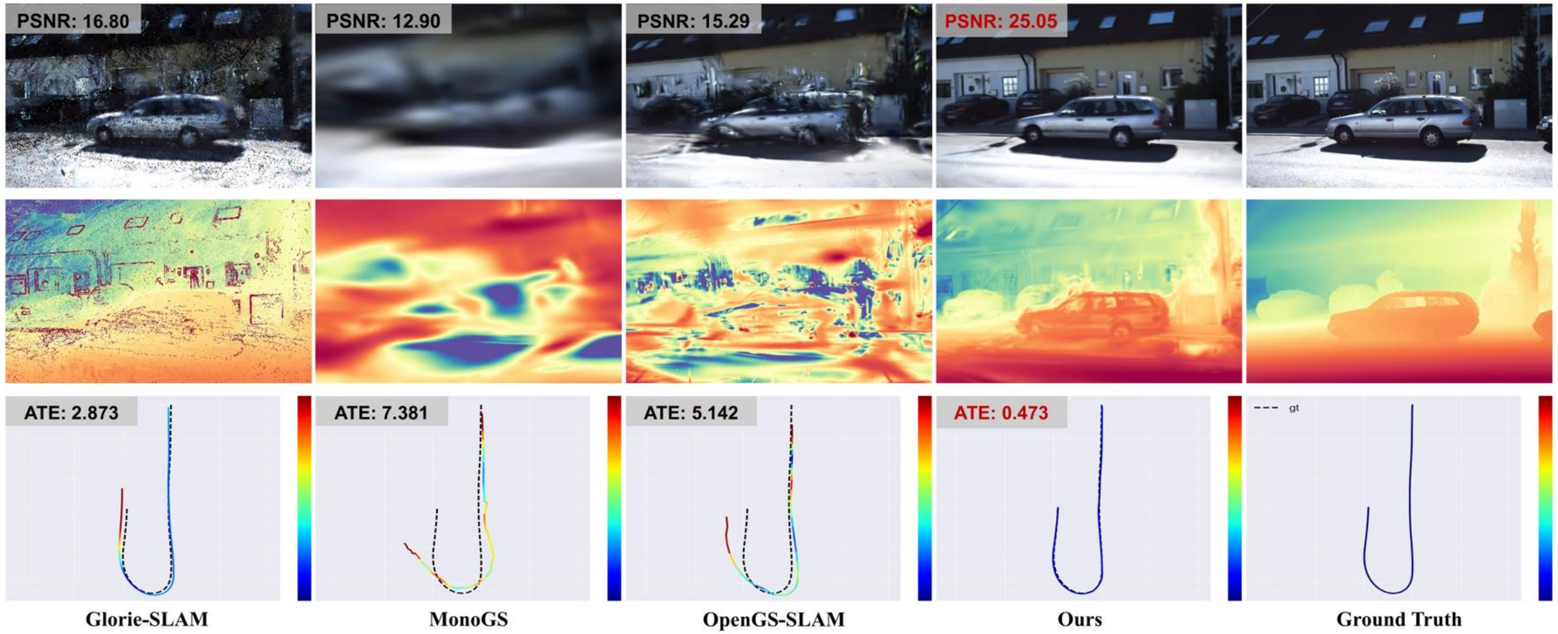
S3PO-GS: Outdoor Monocular SLAM with Global Scale-Consistent 3D Gaussian Pointmaps

Chong Cheng*, Sicheng Yu*, Zijian Wang, Yifan Zhou, Hao Wang
3DAgentWorld Lab, HKUST(GZ)

Motivation

- Outdoor scenes are structurally complex and unbounded, with large viewpoint changes that pose major challenges to pose optimization and reconstruction.
- Purely visual methods lack geometric priors and struggle to converge stably in such environments.
- Existing 3DGS-SLAM methods (e.g., **Photo-SLAM [CVPR'24]**, **MGS-SLAM [RA-L'24]**, **OpenGS-SLAM [ICRA'25]**) rely on separate tracking modules and often suffer from **scale drift** under large motions, degrading localization and mapping accuracy.
- Pretrained monocular depth models (e.g., **DepthAnything**) and point-map models (e.g., **MASt3R**) provide geometric priors, but suffer from **scale inconsistency across frames**, limiting their effectiveness as reliable constraints.

Core: Develop a **3DGS-SLAM framework** that introduces usable **global geometric constraints** under monocular RGB input and suppresses **scale drift**, enabling stable and accurate localization and reconstruction in complex outdoor scenes.



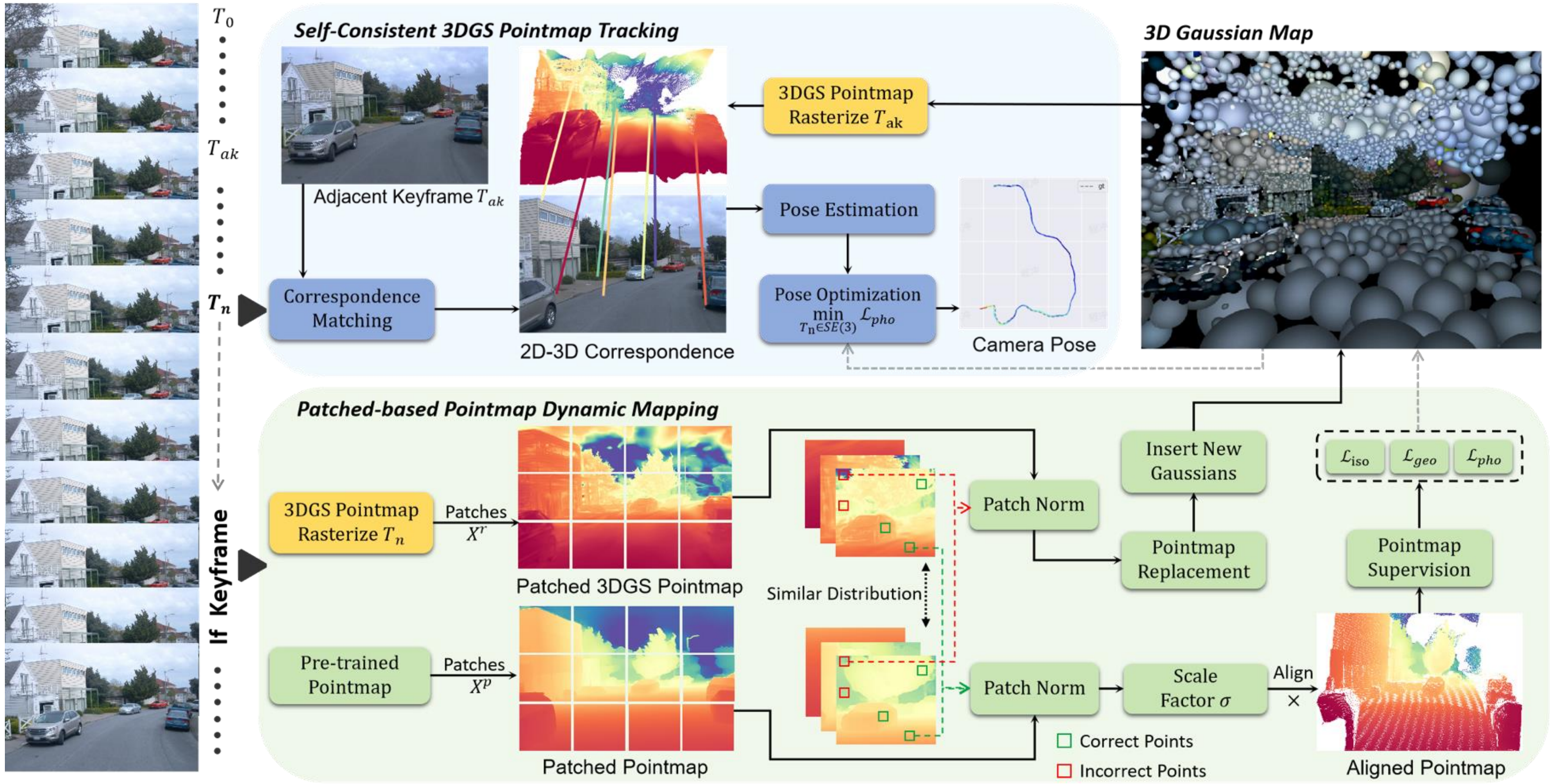
On the **KITTI dataset**, our method maintains **robust camera tracking** and **high-fidelity novel view rendering**, even during sharp turns with large viewpoint changes.

S3PO-GS Framework

The system initializes the 3DGS scene from the first frame and then runs frame by frame with the incoming inputs. Each input frame starts pose tracking, and keyframes start mapping.

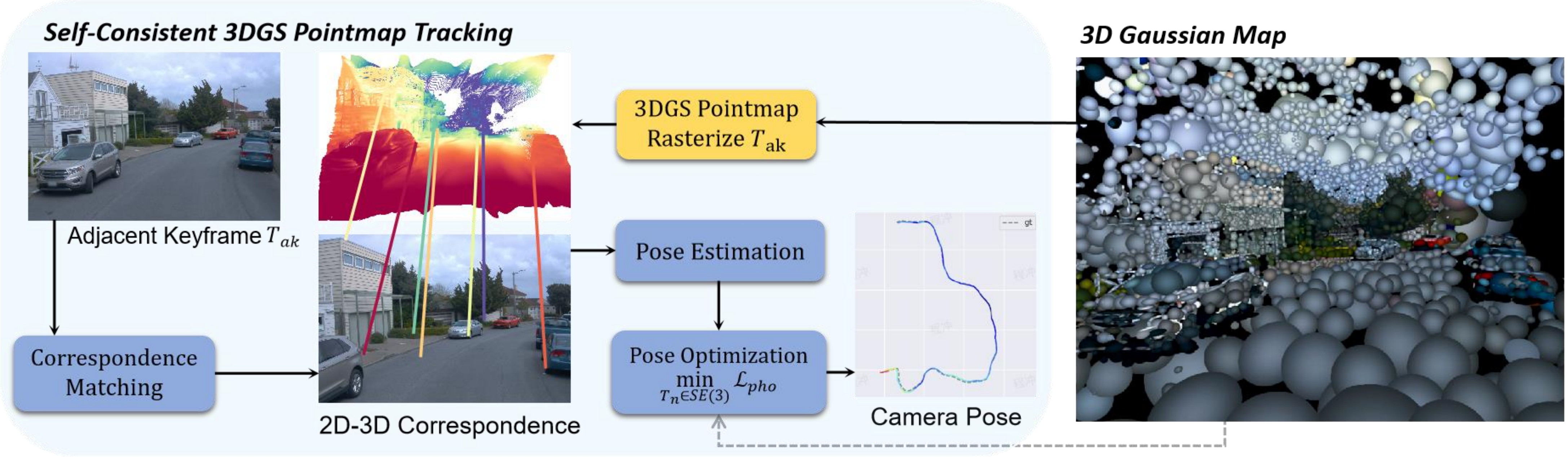
Tracking: Constructs a triple correspondence among 3D Gaussians, point maps, and the new input view for pose estimation, achieving fast convergence of pose optimization while ensuring global scale consistency of the estimated poses.

Mapping: Proposes a patch-based scale alignment module, which performs bidirectional correction between the pretrained point map and the dynamic Gaussian scene, ensuring global scale consistency between poses and the scene.



S3PO-GS Framework

- Starting from the constructed 3DGS scene, render a per-pixel 3DGS point map from the viewpoints of adjacent keyframes.
- Perform per-pixel **2D–2D matching** between the current frame and adjacent keyframes; using pixel correspondences, associate current frame pixels with the rendered point map to obtain **2D–3D correspondences**.
- Based on the **2D–3D correspondences**, estimate the initial pose of the current frame using **PnP + RANSAC**.
- With this initialization, refine the pose by minimizing the photometric error through the **differentiable 3DGS rendering pipeline**.



Self-Consistent 3DGS Pointmap Tracking

3D Gaussian Map

3DGS Pointmap Rasterize T_{ak}

Adjacent Keyframe T_{ak}

Correspondence Matching

2D-3D Correspondence

Pose Estimation

Pose Optimization $\min_{T_n \in SE(3)} \mathcal{L}_{pho}$

Camera Pose

S3PO-GS Framework

Patch-based Scale Alignment

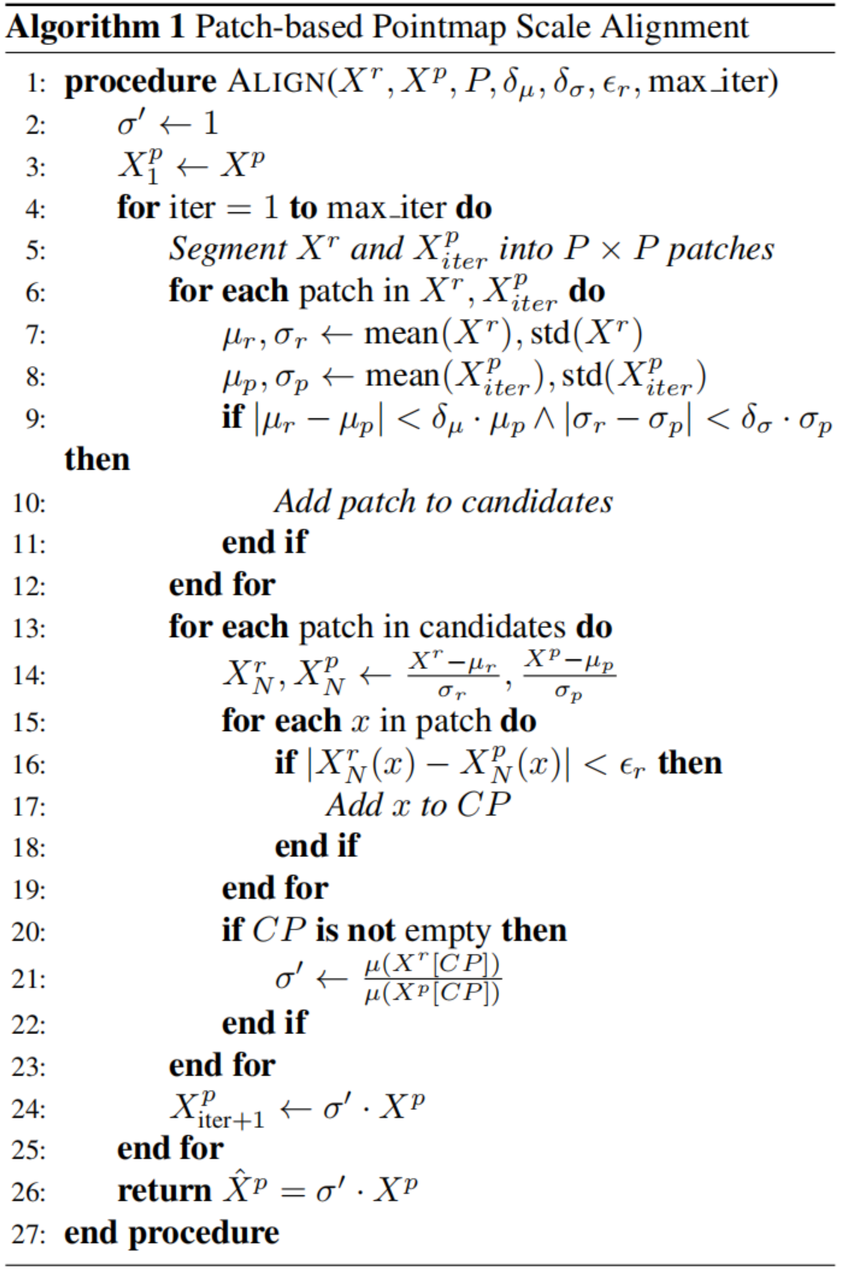
Motivation: Use the geometric prior of the pretrained point map to constrain mapping, but its global scale is inconsistent with the current 3DGS scene and must be corrected first.

Patch-wise alignment: Divide the rendered 3DGS point map and the pretrained point map into patches; first select candidate patches with similar distributions.

Robust estimation: Within candidate patches, normalize and select “accurate pixels” to compute the scale correction factor.

Iterative convergence: Repeat until the scale factor stabilizes.

Result: Obtain a globally scale-aligned pretrained point map, which serves as a reliable geometric constraint to improve mapping consistency and tracking stability.

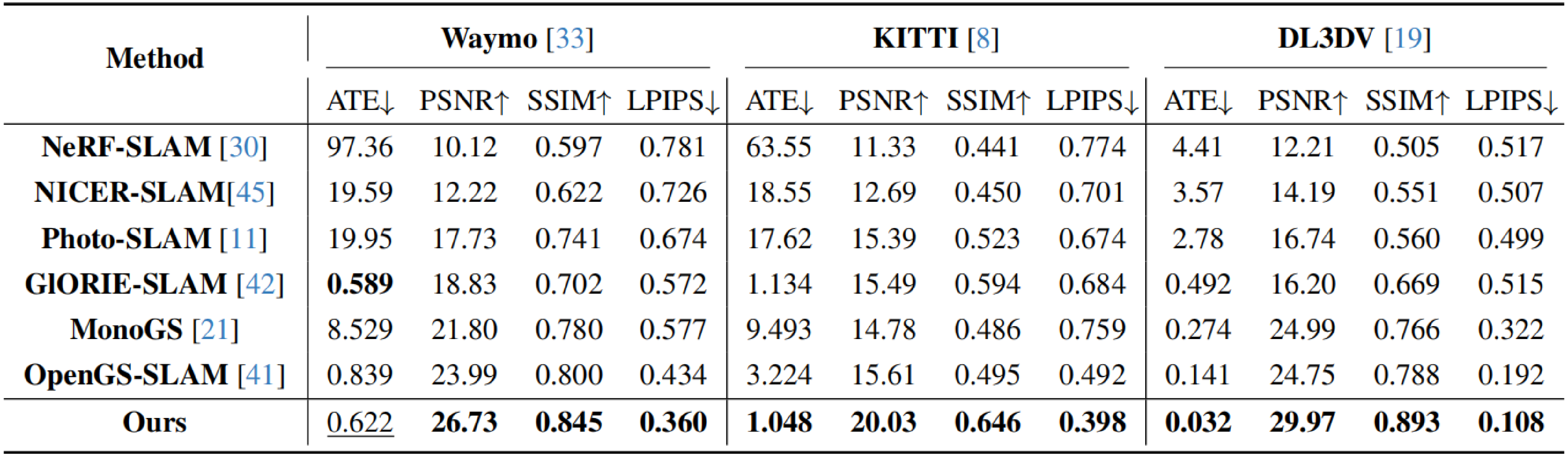


Algorithm 1 Patch-based Pointmap Scale Alignment

```
procedure ALIGN($X^r, X^p, P, \delta_\mu, \delta_\sigma, \epsilon_r$, max_iter)
    $\sigma' \leftarrow 1$
    $X_1^p \leftarrow X^p$
    for iter $= 1$ to max_iter do
        Segment $X^r$ and $X^p_{iter}$ into $P \times P$ patches
        for each patch in $X^r, X^p_{iter}$ do
            $\mu_r, \sigma_r \leftarrow \text{mean}(X^r), \text{std}(X^r)$
            $\mu_p, \sigma_p \leftarrow \text{mean}(X^p_{iter}), \text{std}(X^p_{iter})$
            if $|\mu_r - \mu_p| < \delta_\mu \cdot \mu_p \wedge |\sigma_r - \sigma_p| < \delta_\sigma \cdot \sigma_p$ then
                Add patch to candidates
            end if
        end for
        for each patch in candidates do
            $X^r_N, X^p_N \leftarrow \frac{X^r - \mu_r}{\sigma_r}, \frac{X^p - \mu_p}{\sigma_p}$
            for each $x$ in patch do
                if $|X^r_N(x) - X^p_N(x)| < \epsilon_r$ then
                    Add $x$ to $CP$
                end if
            end for
            if $CP$ is not empty then
                $\sigma' \leftarrow \frac{\mu(X^r[CP])}{\mu(X^p[CP])}$
            end if
        end for
        $X^p_{\text{iter}+1} \leftarrow \sigma' \cdot X^p$
    end for
    return $\hat{X}^p = \sigma' \cdot X^p$
end procedure
```

Results

On three outdoor datasets, our method achieves **significant performance improvements** compared to previous **3DGS-based** and **NeRF-based** SLAM methods.



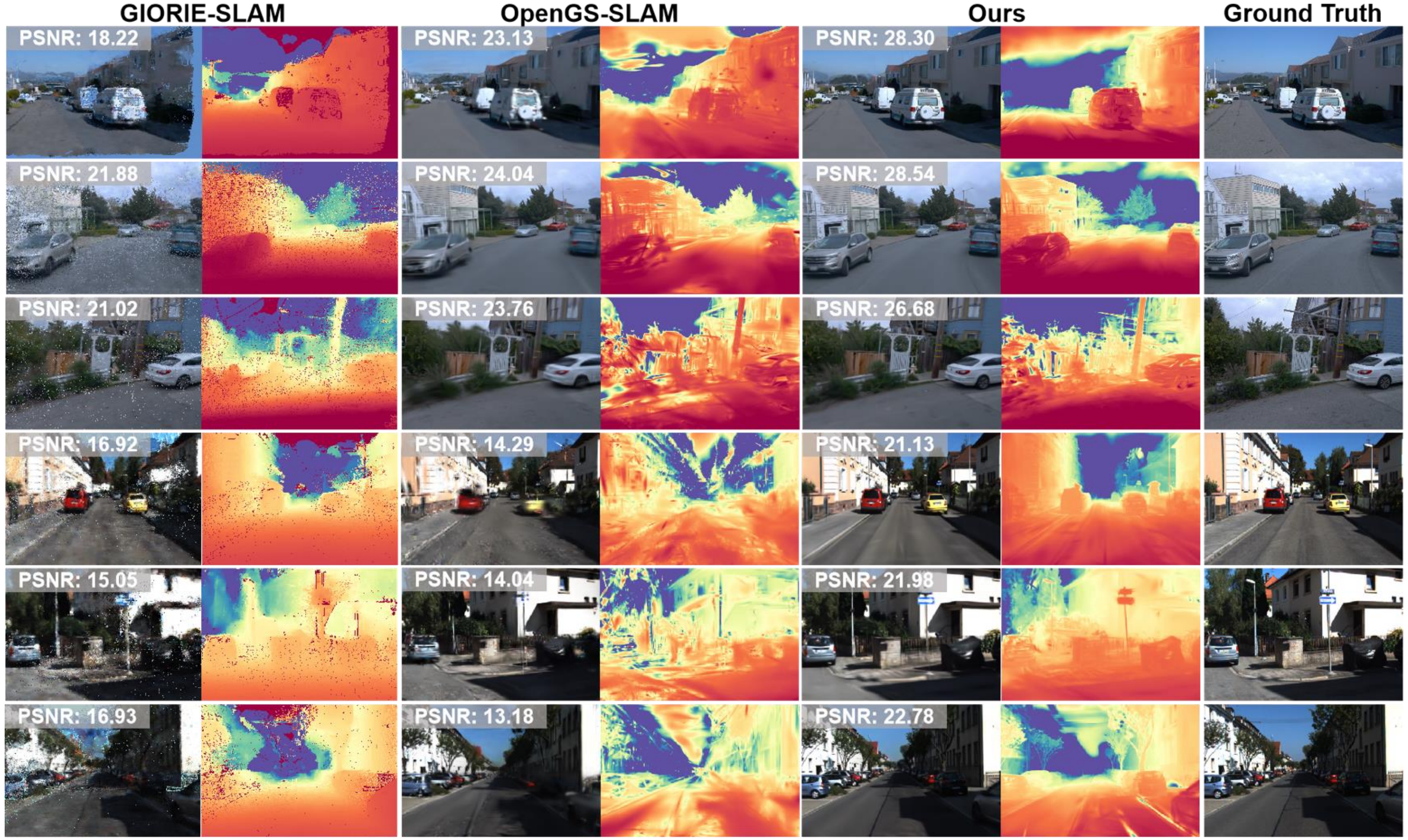
Method	Waymo [33]				KITTI [8]				DL3DV [19]			
	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓	ATE↓	PSNR↑	SSIM↑	LPIPS↓
NeRF-SLAM [30]	97.36	10.12	0.597	0.781	63.55	11.33	0.441	0.774	4.41	12.21	0.505	0.517
NICER-SLAM[45]	19.59	12.22	0.622	0.726	18.55	12.69	0.450	0.701	3.57	14.19	0.551	0.507
Photo-SLAM [11]	19.95	17.73	0.741	0.674	17.62	15.39	0.523	0.674	2.78	16.74	0.560	0.499
GlORIE-SLAM [42]	**0.589**	18.83	0.702	0.572	1.134	15.49	0.594	0.684	0.492	16.20	0.669	0.515
MonoGS [21]	8.529	21.80	0.780	0.577	9.493	14.78	0.486	0.759	0.274	24.99	0.766	0.322
OpenGS-SLAM [41]	0.839	23.99	0.800	0.434	3.224	15.61	0.495	0.492	0.141	24.75	0.788	0.192
Ours	0.622	**26.73**	**0.845**	**0.360**	**1.048**	**20.03**	**0.646**	**0.398**	**0.032**	**29.97**	**0.893**	**0.108**

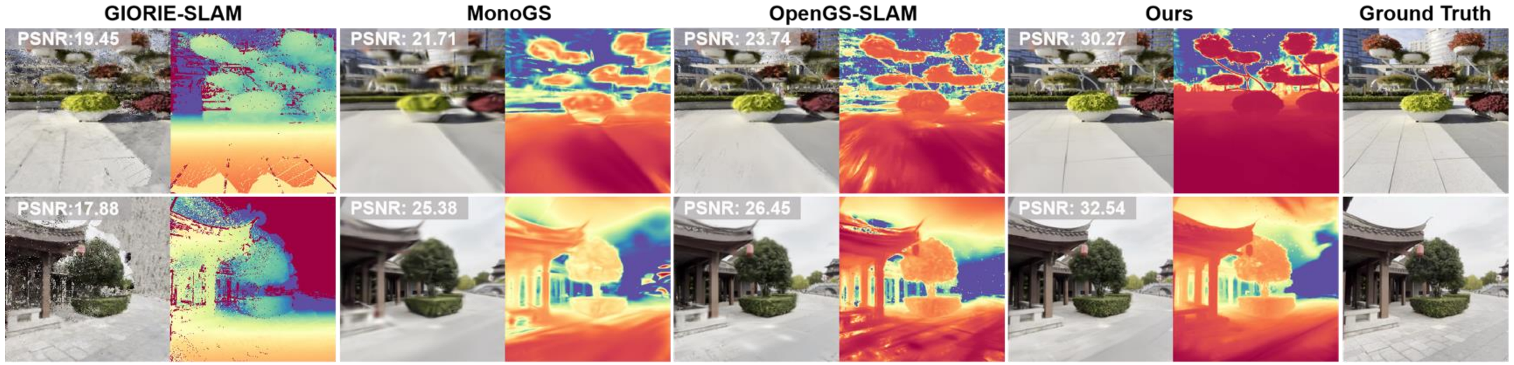
Results

The right figure shows novel view rendering comparisons on **Waymo (rows 1–3)**, **KITTI (rows 4–6)**, and **DL3DV (last two rows)**, including both RGB images and depth maps.

It can be observed that:

- Our method renders **high-quality images** across diverse outdoor scenes, with higher fidelity in details such as vehicles, streets, buildings, flowerbeds, and eaves.
- The depth maps are more accurate in complex regions like tree branches, roadside vehicles, and pillars, and show greater stability and smoothness in large areas such as the ground.



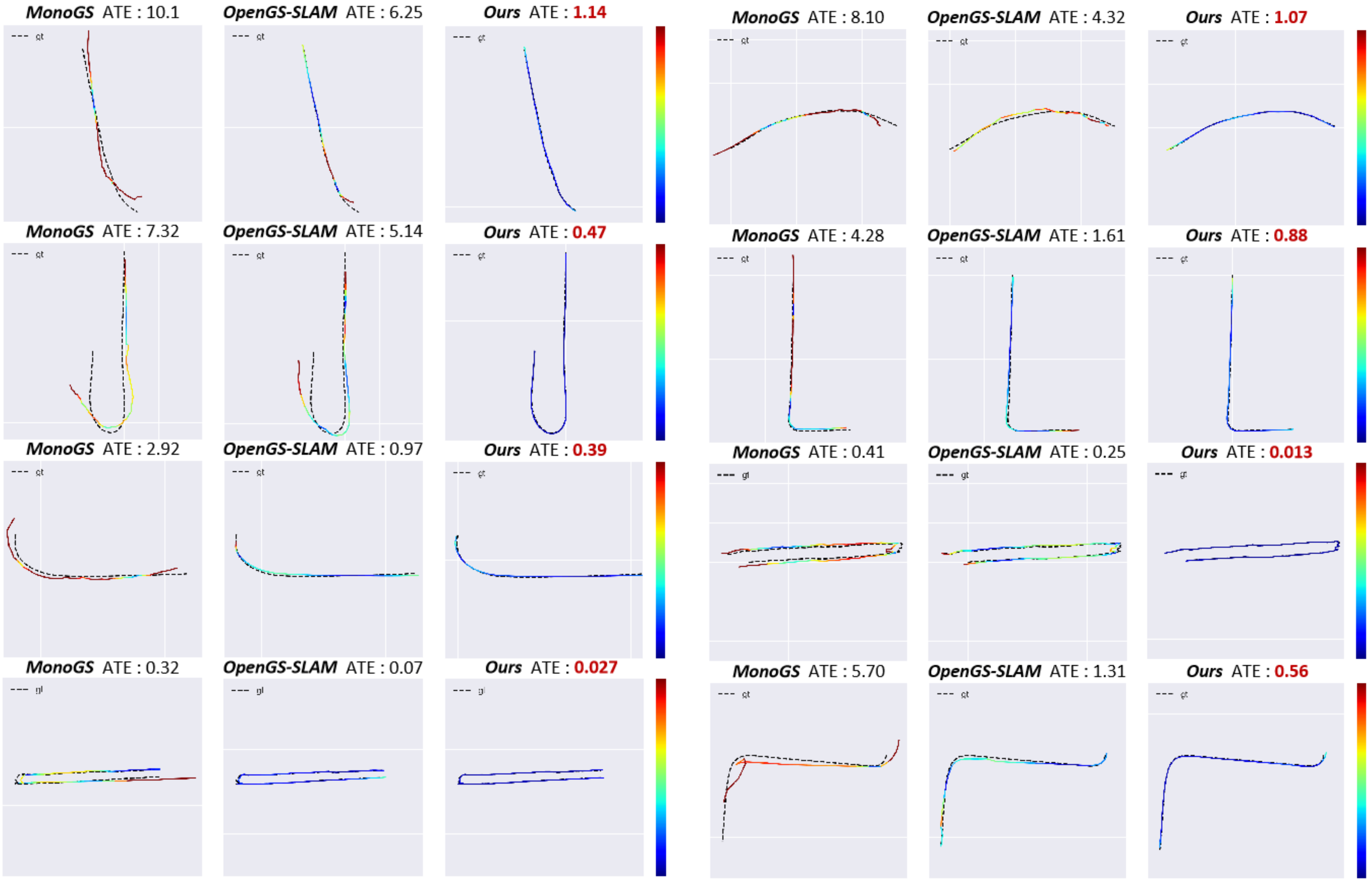


Results

The right figure shows pose trajectory comparisons with other 3DGS-SLAM methods across multiple scenes from three datasets.

It can be observed that:

- Under large viewpoint changes, other methods exhibit significant drift, while our method demonstrates superior robustness.



Pose trajectory comparison (meter)

Results Demo

The video demonstrates the real-time performance of our SLAM system.

While processing the input video stream, the system simultaneously optimizes the camera trajectory and the 3D Gaussian scene.

- **Green box**: Camera pose of the current input frame.
- **Blue box**: Camera poses of historical keyframes.

THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

Thank You