PVDepth: Panoramic Video Depth Estimation via Geometry-aware Spatiotemporal Adaptation

Chuanxin Song, Peixi Peng (Peking University & Pengcheng Laboratory)

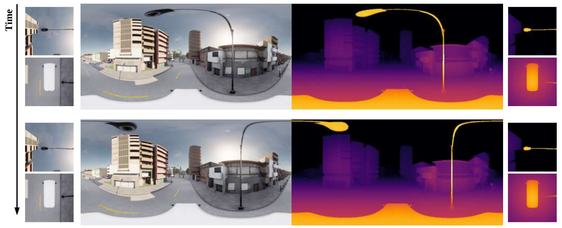
ICML International Conference On Machine Learning

TL;DR: End-to-end panoramic video depth estimation with a large-scale RGB-D dataset and geometry-aware spatiotemporal adaptation.

Challenges in Panoramic Video Depth Estimation

Data scarcity: there is no public large-scale panoramic RGB-D video benchmark.

Panoramic-perspective domain gap:
ERP-specific spatial distortion
Non-linear temporal dynamics



ERP stretches polar regions and induces irregular temporal dynamics.

Why Existing Methods Limited

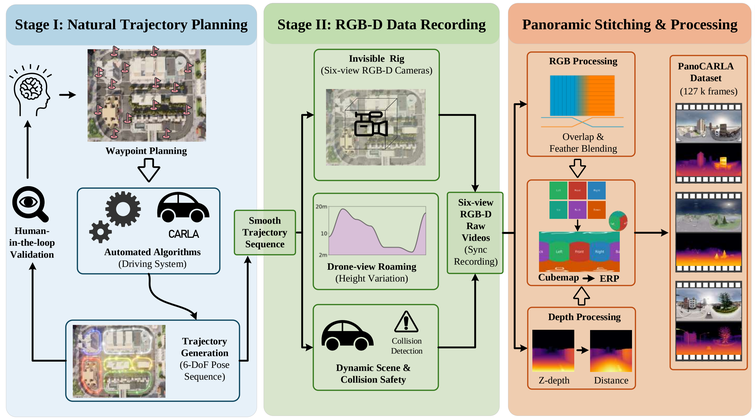
Panoramic image-based methods: good per-frame results, bad temporal consistency.

Post-processing alignment methods: computationally expensive and struggles with dynamic scenes.

Perspective video depth methods: sub-optimal adaptation suffer from spatiotemporal domain gap.

Need: training data && panoramic geometry-aware adaptation

PanoCARLA Dataset



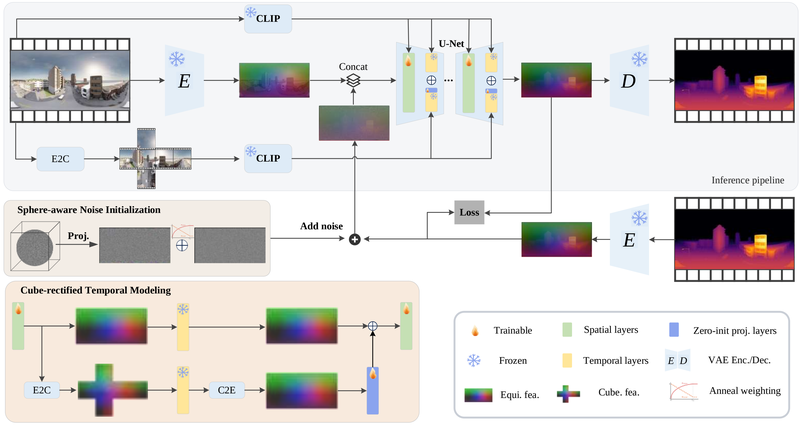
With self-occlusion | Without self-occlusion

- RGB, depth, and camera poses.
- 50.61 km drone-view roaming.

Dataset	Structure -3D	Matterport -3D	Stanford -2D3D	Omni360 -Scene	PanoCARLA
Is Video?	×	×	×	×	✓
#Frames	196K	10K	1K	61K	**126.7K**

Data scarcity → **PanoCARLA** | **Spatial distortion** → **PSNI** | **Temporal nonlinearity** → **CRTM**

PVDepth Framework



Built upon DepthCrafter.

PSNI (*Progressive Sphere-aware Noise Initialization strategy*)**:**
Anneal noise from planar to spherical.
Adapt to non-uniform ERP density.

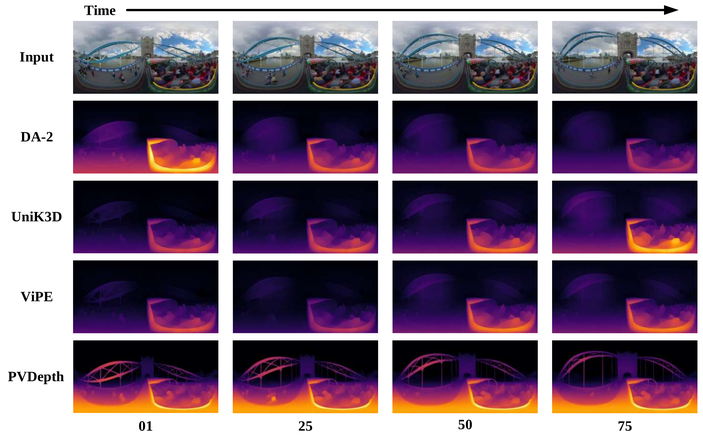
CRTM (*Cube-rectified Temporal Modeling module*):
An auxiliary cubemap temporal branch.
Rectify non-linear temporal dynamics.

Experiments

Quantitative comparison (avg over all settings)

Method	AbsRel ↓	δ_1 ↑
Panorama image depth-based		
DA-2$_{ZS}$ (Li et al., 2025)	1.171 (0.757)	0.615 (0.754)
UniK3D$_{ZS}$ (Piccinelli et al., 2025)	0.592 (0.461)	0.638 (0.736)
PanoDA$_{ZS}$ (Cao et al., 2025)	0.710 (0.526)	0.280 (0.342)
DepthAnywhere$_{ZS}$ (Wang & Liu, 2024)	0.449 (0.233)	0.464 (0.800)
DepthAnywhere$_{FT}$ (Wang & Liu, 2024)	0.458 (**0.207**)	0.483 (**0.846**)
Post-processing alignment-based		
Matrix-3D (Yang et al., 2025b)	0.494	0.513
ViPE (UniK3D-based) (Huang et al., 2025)	0.670	0.721
Perspective video depth-based		
VideoDepthAnything$_{ZS}$ (Chen et al., 2025)	0.356	0.505
DepthAnyVideo$_{ZS}$ (Yang et al., 2025a)	0.295	0.608
DepthCrafter(Cube-based)$_{ZS}$	0.941	0.328
DepthCrafter(Cube-based)$_{FT}$	0.417	0.590
DepthCrafter$_{ZS}$ (Hu et al., 2025)	0.290	0.567
DepthCrafter$_{FT}$ (Hu et al., 2025)	0.232	0.754
PVDepth (Ours)	**0.199**	**0.798**

Qualitative comparison on In-the-Wild Videos



PVDepth achieves higher accuracy and temporal stability.

Ablation study (avg over all settings)

Method	PSNI	CRTM	AbsRel ↓	δ_1 ↑
Baseline	–	–	0.248	0.757
w/ PSNI	✓	–	0.228	0.782
w/ CRTM	–	✓	0.208	0.779
PVDepth (Full)	✓	✓	**0.199**	**0.798**