CVPR JUNE 3-7, 2026 DENVER COLORADO

VGA: Empowering Aerial-Ground Localization by Visual Geometry Alignment

Tao Jun Lin[1], Yujiao Shi[2], Hongdong Li[1,3]

[1]The Australian National University, [2]ShanghaiTech University, [3]Amazon

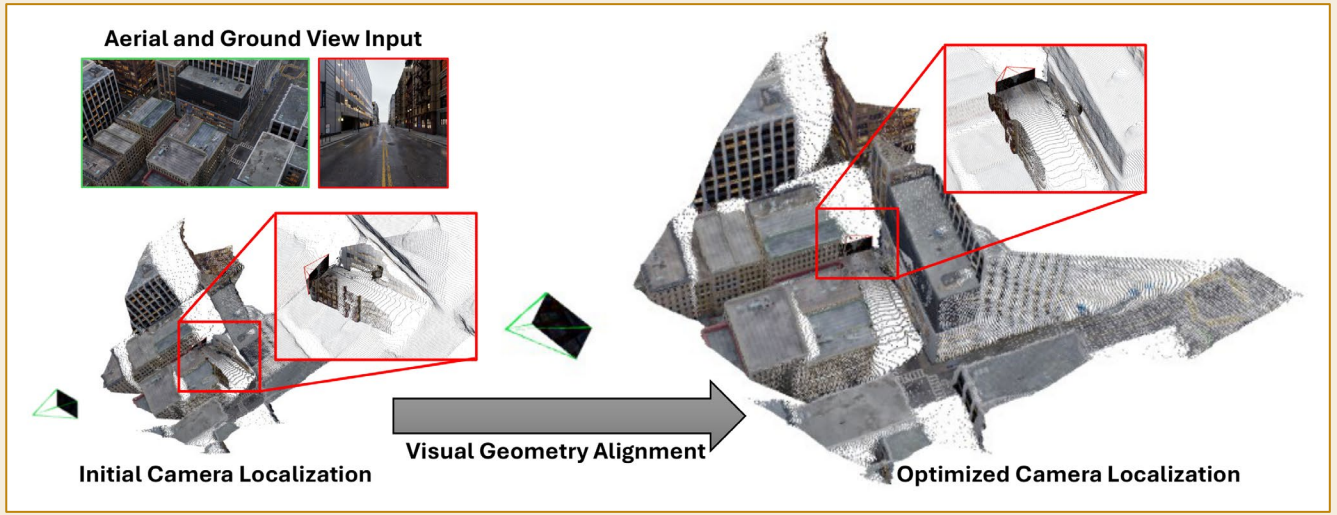
Australian National University

上海科技大学 ShanghaiTech University

amazon

INTRODUCTION

We explore the practical benefits of regularizing 6-DoF aerial-ground visual localization with physically grounded priors estimated from visual input.



Problem Setup: Given uncalibrated aerial and ground pair, we aim to estimate accurate 6-DoF relative camera pose between views.

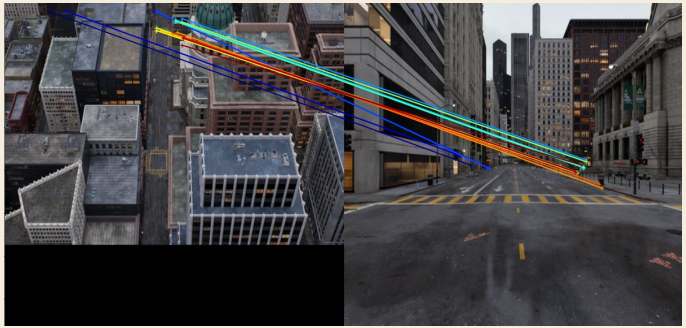
MOTIVATION

Aerial–Ground Localization is Fundamentally Hard.

CHALLENGES

1. **Extreme viewpoint change**

2. **Existing methods fail**



Example Matches by MASt3R

CONTRIBUTIONS

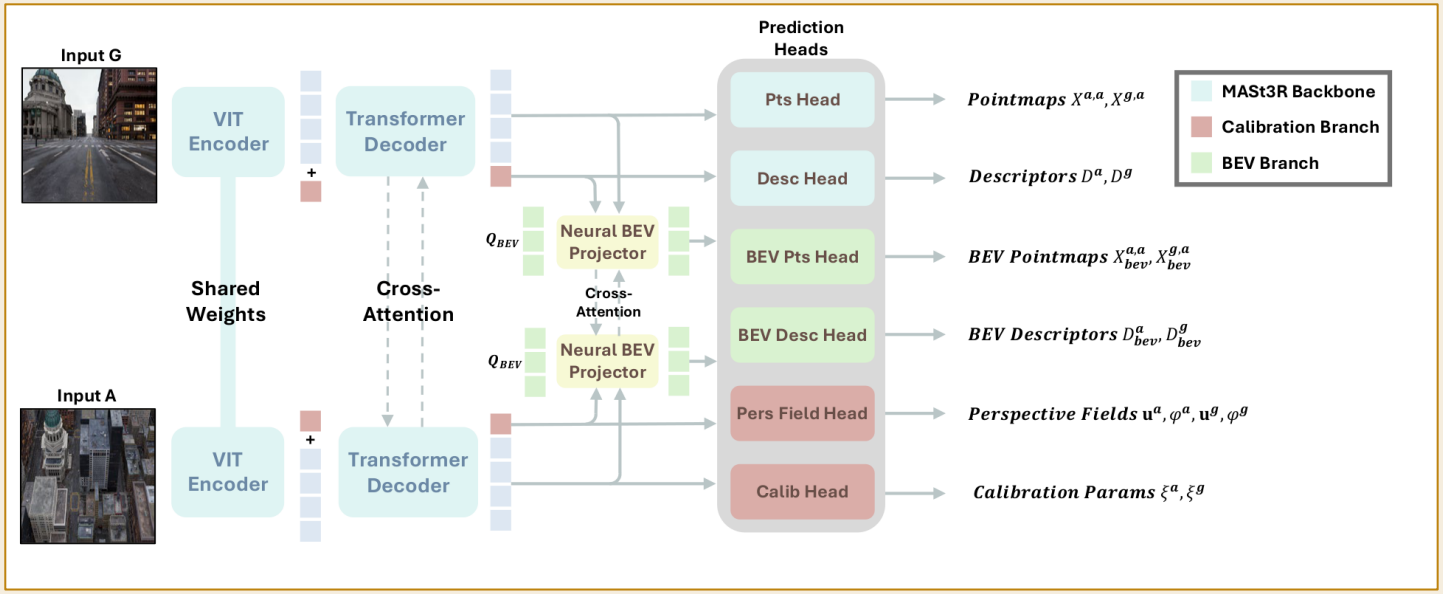
1. **Unified Dense Geometry Feed-Forward Framework**

2. **Inference-Time Refinement with Physically Grounded Priors**

3. **SOTA Performance**

FRAMEWORK

Stage 1: Dense Geometry Regression

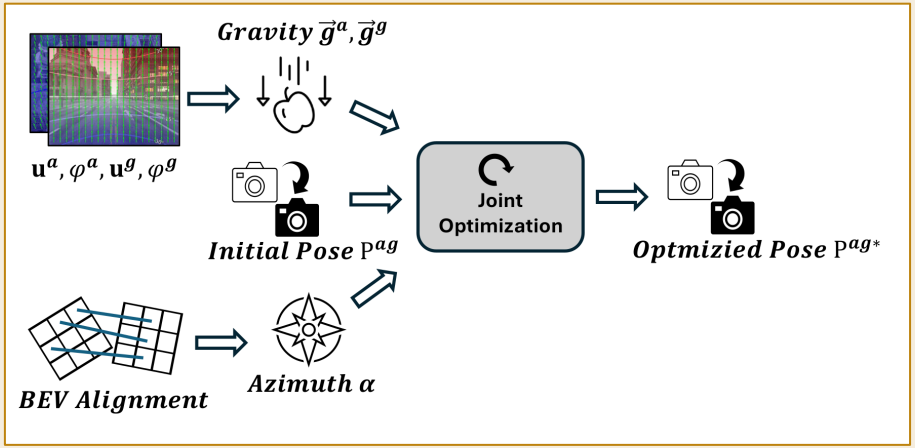


We augment existing Siamese ViT backbone with:

- **Calibration Branch** for dense perspective field and camera parameter prediction;
- **BEV prediction branch** for canonical BEV geometry and descriptor prediction.

INFERENCE

Stage2: Joint Pose–Geometry Optimization



$$\underset{\mathrm{P}^{ag}}{\operatorname{argmin}} \boldsymbol{\lambda}_S E_S + \boldsymbol{\lambda}_g E_g + \boldsymbol{\lambda}_\alpha E_\alpha$$

Sampson reprojection Error

Perspective-space inlier correspondences.

$$E_S = \sum_{i=1}^{N} \frac{|X_i^{g\,T} E\, X_i^a|^2}{\left\|E_{12} X_i^a\right\|^2 + \left\|(E^T)_{12} X_i^a\right\|^2}$$

Gravity alignment

Regularizes the roll–pitch components of rotation.

$$E_g = \left\| R^{ag} \overrightarrow{\mathbf{g}^g} - \overrightarrow{\mathbf{g}^a} \right\|_2^2$$

Azimuth constraint

Planar yaw α from BEV Procrustes alignment.

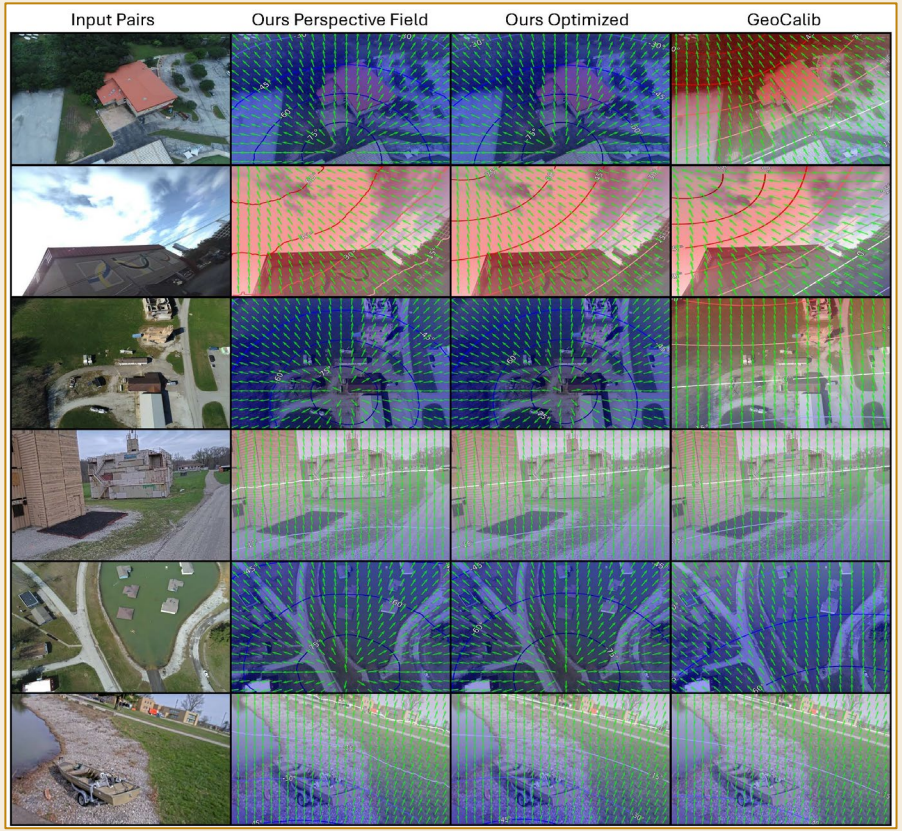
$$E_\alpha = \| \log\left(R_z^{ag\,T} R_z(\alpha) \right) \|_2^2$$

Adds only a few milliseconds of overhead at inference.

GRAVITY PRIOR

Gravity-Guided Calibration via Perspective Fields

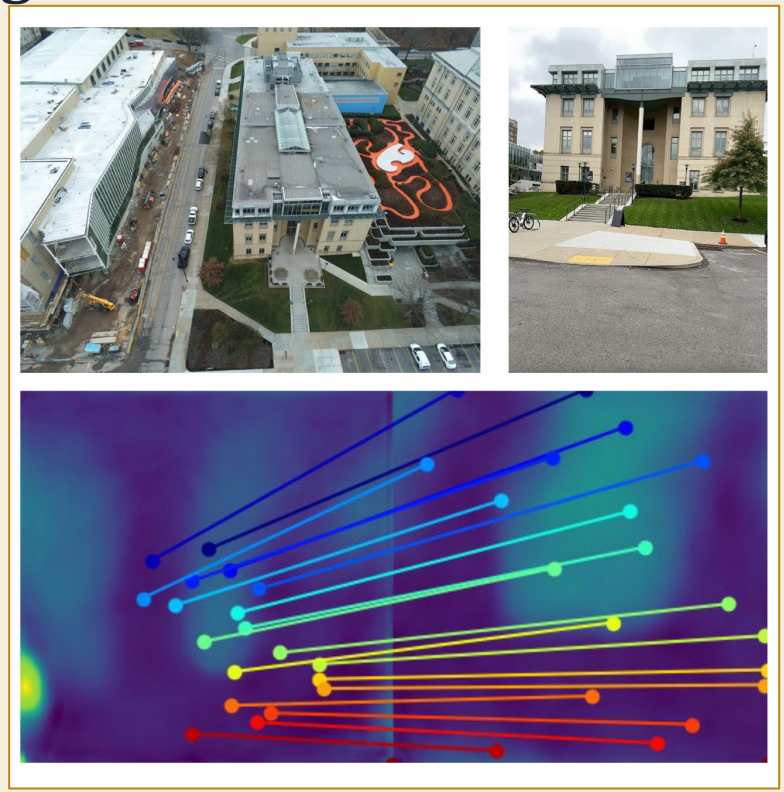
- Calibration head predicts per-view vertical FoV, principal point, and pitch & roll relative to gravity.
- A dense perspective field (per-pixel up-vectors + latitude) is learned from vertical structures and ground planes.
- Our perspective field prediction outperforms GeoCalib, yielding more accurate gravity estimation for aerial-ground localization.



PLANAR PRIOR

Canonical Bird's-Eye-View Alignment

- A Neural BEV Projector lifts both views into a shared, gravity-aligned metric BEV plane.
- Performs stable 4-DoF planar alignment (azimuth, xy, scale) in BEV Plane.
- BEV pointmaps and descriptors give azimuth α via Procrustes Alignment (Kabsch–Umeyama + RANSAC).



Predicted BEV geometry and cross-view BEV correspondences

Experiments

State-of-the-Art Aerial–Ground Localization

	RTA@2°	RTA@5°	RTA@15°	RTA@25°	RRA@2°	RRA@5°	RRA@15°	RRA@25°	AUC@30°
LoFTR [62]	0.10	0.32	2.16	6.28	0.04	0.08	0.34	0.74	0.28
SP+SG [48]	0.04	0.16	1.52	3.66	0.02	0.06	0.16	0.46	0.16
ROMA [14]	2.42	3.16	6.00	10.82	2.36	2.86	3.68	4.60	3.59
MASt3R (Released) [26]	-	-	-	-	-	-	-	-	-
VGGT [72]	0.08	0.40	4.46	11.74	0.02	0.12	0.76	1.96	1.66
π^3 [78]	0.04	0.26	4.12	17.48	0.40	2.56	9.64	16.00	1.52
MASt3R (AerialMegaDepth) [69]	6.72	8.50	15.50	24.92	5.86	7.98	9.26	13.52	9.78
MASt3R (AerialMegaDepth+MatrixCity)	15.30	26.30	32.88	37.10	13.06	25.70	31.98	33.54	29.12
VGA (Ours)	**17.32**	**31.40**	**39.76**	**45.64**	**16.02**	**32.64**	**42.70**	**46.70**	**34.97**

Quantitative Results on MatrixCity(BigCity)

	RTA@2°	RTA@5°	RTA@15°	RTA@25°	RRA@2°	RRA@5°	RRA@15°	RRA@25°	AUC@30°
LoFTR [62]	0.00	0.06	0.78	1.88	0.04	0.48	0.88	1.88	0.36
SP+SG [48]	0.06	0.56	1.98	3.92	0.04	0.48	1.98	2.84	1.22
ROMA [14]	0.16	1.58	38.02	43.26	5.28	24.28	40.24	43.26	26.93
MASt3R (AerialMegaDepth) [69]	0.66	5.56	48.56	54.94	5.24	28.64	49.88	52.90	34.75
VGGT [72]	0.44	4.20	33.32	40.82	4.52	27.40	50.60	54.72	25.13
π^3 [78]	**1.28**	**8.30**	57.24	71.26	6.94	36.88	72.16	75.70	43.45
VGA (Ours)	0.24	5.86	**65.18**	**72.74**	**8.94**	**42.36**	**72.72**	**76.34**	**47.45**

Quantitative Results on ACC-NVS1 and ULTRRA

	RTA@2°	RTA@5°	RTA@15°	RTA@25°	RRA@2°	RRA@5°	RRA@15°	RRA@25°	AUC@30°
Baseline	15.30	26.30	32.88	37.10	13.06	25.70	31.98	33.54	29.12
+ Planar Alignment	16.24	27.66	33.64	37.92	13.68	27.34	33.58	35.62	31.36
+ Gravity Prior	**17.70**	30.76	36.38	40.60	15.72	31.56	38.34	41.28	33.38
Joint Optimization	17.32	**31.40**	**39.76**	**45.64**	**16.02**	**32.64**	**42.70**	**46.70**	**34.97**

Ablation of proposed geometric priors

~11%

AUC@30° gain vs. next-best

29.1 → 35.0

MatrixCity AUC@30° (baseline → VGA)

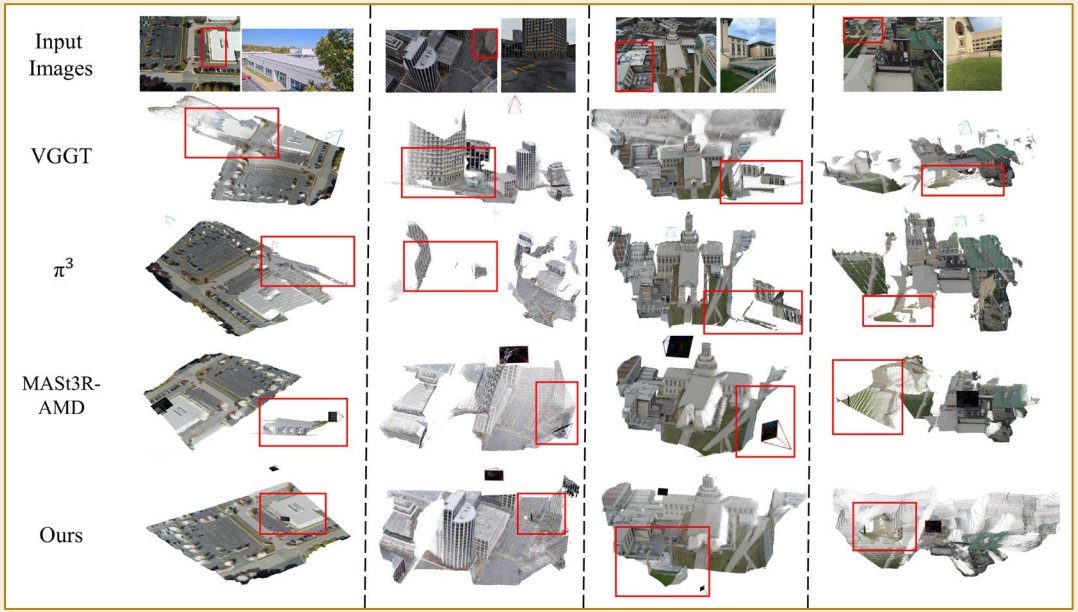
47.45

ULTRRA / ACC-NVS1 — SOTA zero-shot

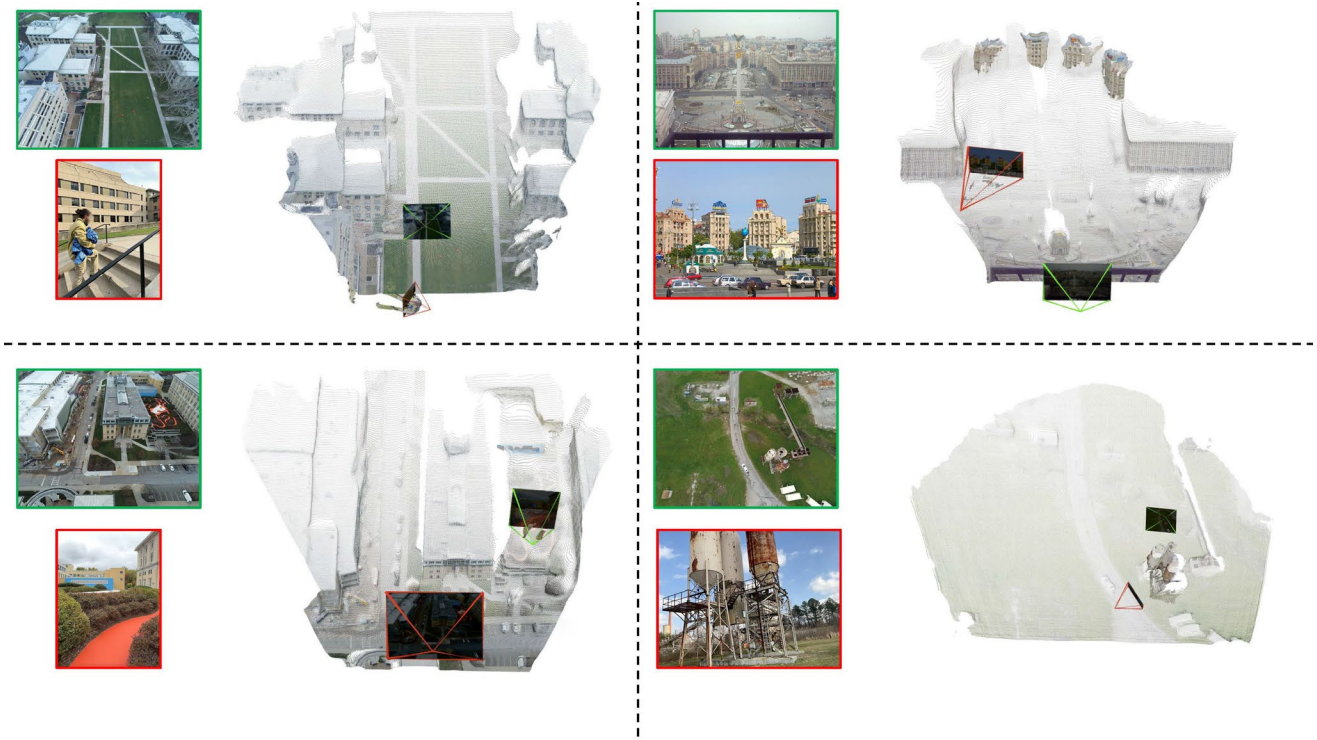
Experiments

VGA Localizes Where Others Fail

- Compared with VGGT, π^3 and MASt3R-AMD.
- Red boxes mark the recovered ground-camera pose.
- VGA reconstructs coherent shared geometry under extreme baselines and altitude.



THANK YOU



Australian National University