UCF | Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

Georeferenced Cross-View 3D Reconstruction from Ground and Satellite Images

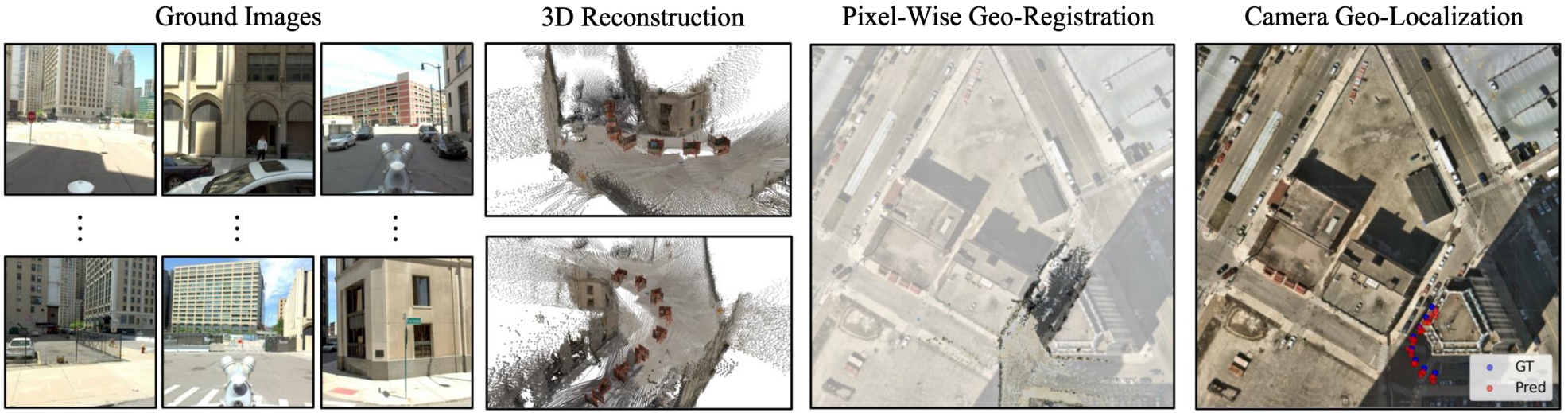
David G. Shatwell, Mubarak Shah

*Accepted in **NeurIPS 2026***

Cross-View 3D Reconstruction

Given a set of N ground images and one georeferenced satellite image:

- Create a metric **3D point-cloud** of the scene
- Estimate the GPS coordinate of every pixel (**geo-registration**)
- Estimate the GPS coordinate of every camera (**geo-localization**)



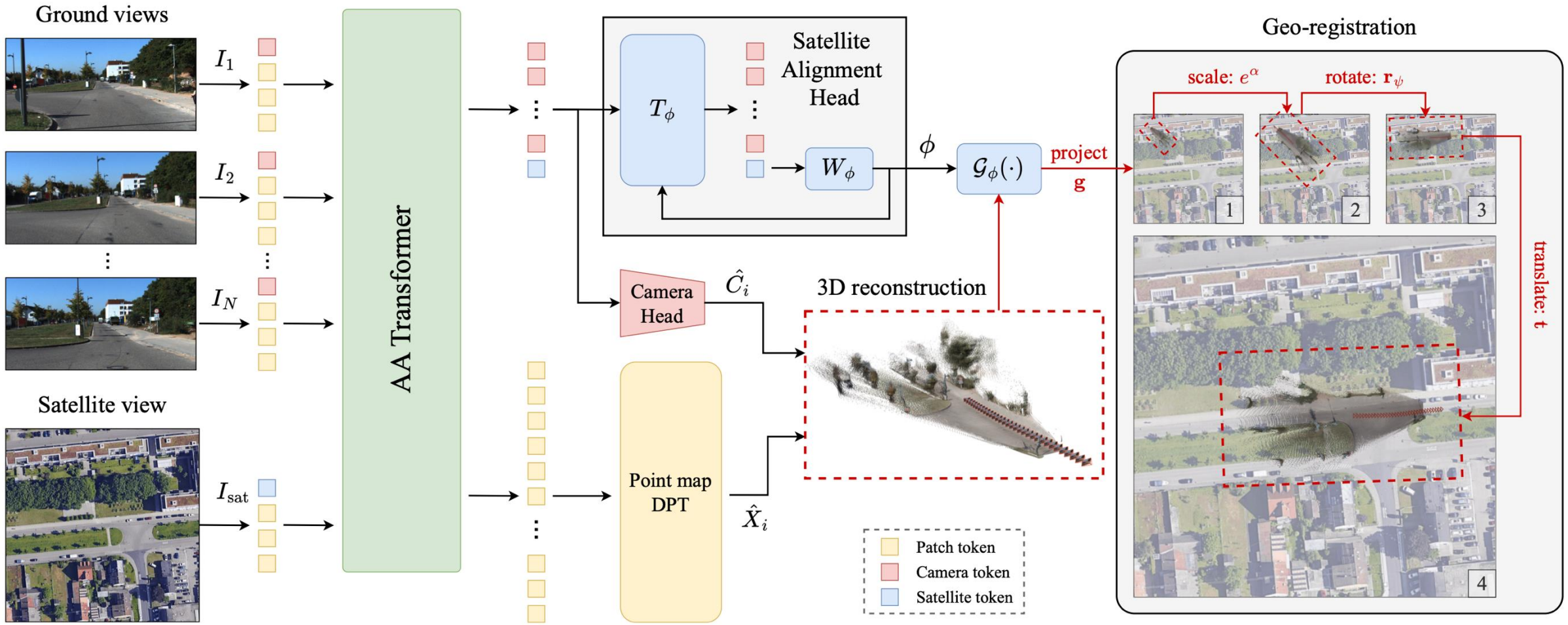
Proposed Solution

X-VGGT: Cross(**X**)-**V**iew **G**eometry **G**rounded **T**ransformer

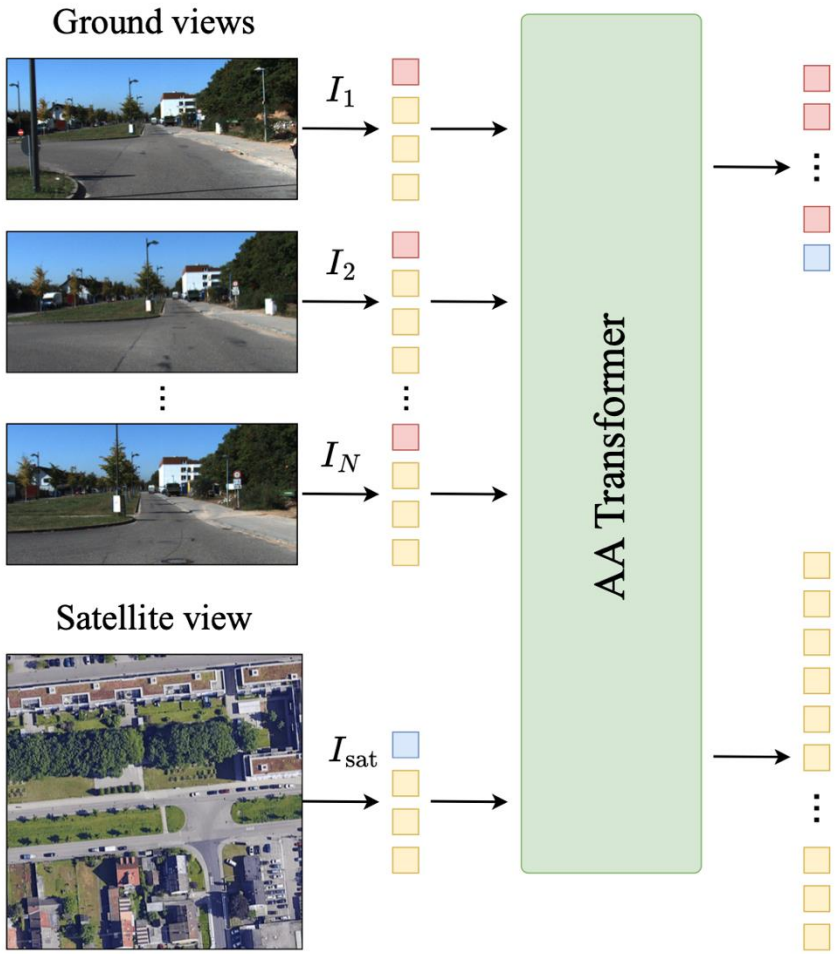
Geo-registration between satellite and *N* ground views:

1. Predict **3D reconstruction** (point-maps and camera poses) for each ground view
2. Predicts $\mathcal{G}_\phi: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ between 3D reconstruction and orthographic satellite view
   - $\mathcal{G}_\phi$ parameterized by gravity ($\mathbf{g}$), rotation (θ), translation ($\mathbf{t}$) and scale (α)
3. Transforms 3D points and camera centers into satellite reference frame
4. Assign GPS coordinate to every ground pixel and camera

Architecture

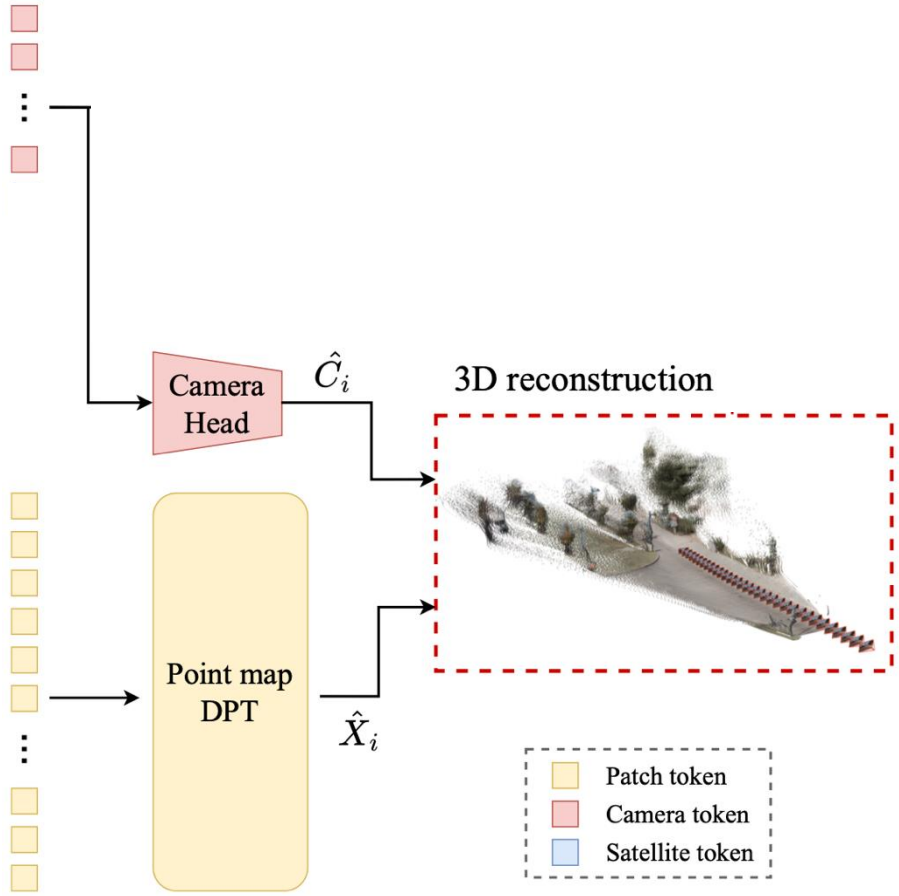


Architecture: **Backbone**



- Tokenize ground $(I_1, \ldots, I_N)$ and satellite (I_{sat}) views with DINOv2
- Append learnable camera and satellite tokens
- Alternating Attention Transformer:
  - Global attention
  - Frame attention
- Output: geometrically refined tokens

Architecture: **3D Reconstruction of Ground Views**



Camera Parameters

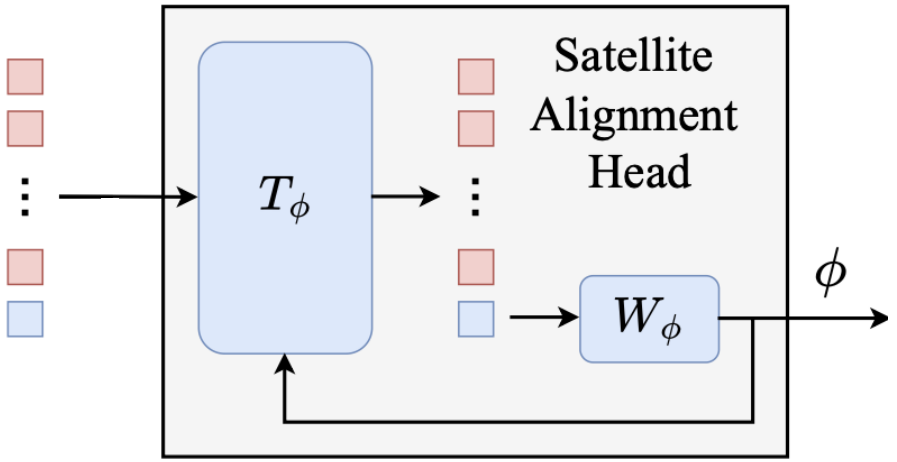
- Camera tokens are decoded using a Camera Head
- Each camera $\hat{C}_i$ is represented by an 8-vector:
  - Translation: $\mathbf{t}_i \in \mathbb{R}^3$
  - Rotation: $\mathbf{q}_i \in \mathbb{R}^4$, s.t. $\|q_i\| = 1$
  - Focal length: $f \in \mathbb{R}$
- First camera ($\hat{C}_0$) defines the reference frame

Point Maps

- Point maps are decoded with (DPT)
- 3D point & confidence for each pixel
- Represented by $\hat{X}_i \in \mathbb{R}^{H \times W \times 3}$

Architecture: **Satellite Alignment Head**

- Predicts transformation parameters ϕ between ground and satellite reference frames
- Satellite and ground camera tokens refined with lightweight transformer T_ϕ
- Linear projection W_ϕ maps refined satellite token to 8-vector ϕ
- ϕ parameterized by:
  - Gravity: $\mathbf{g} \in \mathbb{R}^3$, s.t. $\|\mathbf{g}\| = 1$
  - Metric scale: $\alpha \in \mathbb{R}$
  - Rotation: $\mathbf{r}_\psi = (\cos\psi, \sin\psi)$
  - Translation: $\mathbf{t} \in \mathbb{R}^2$

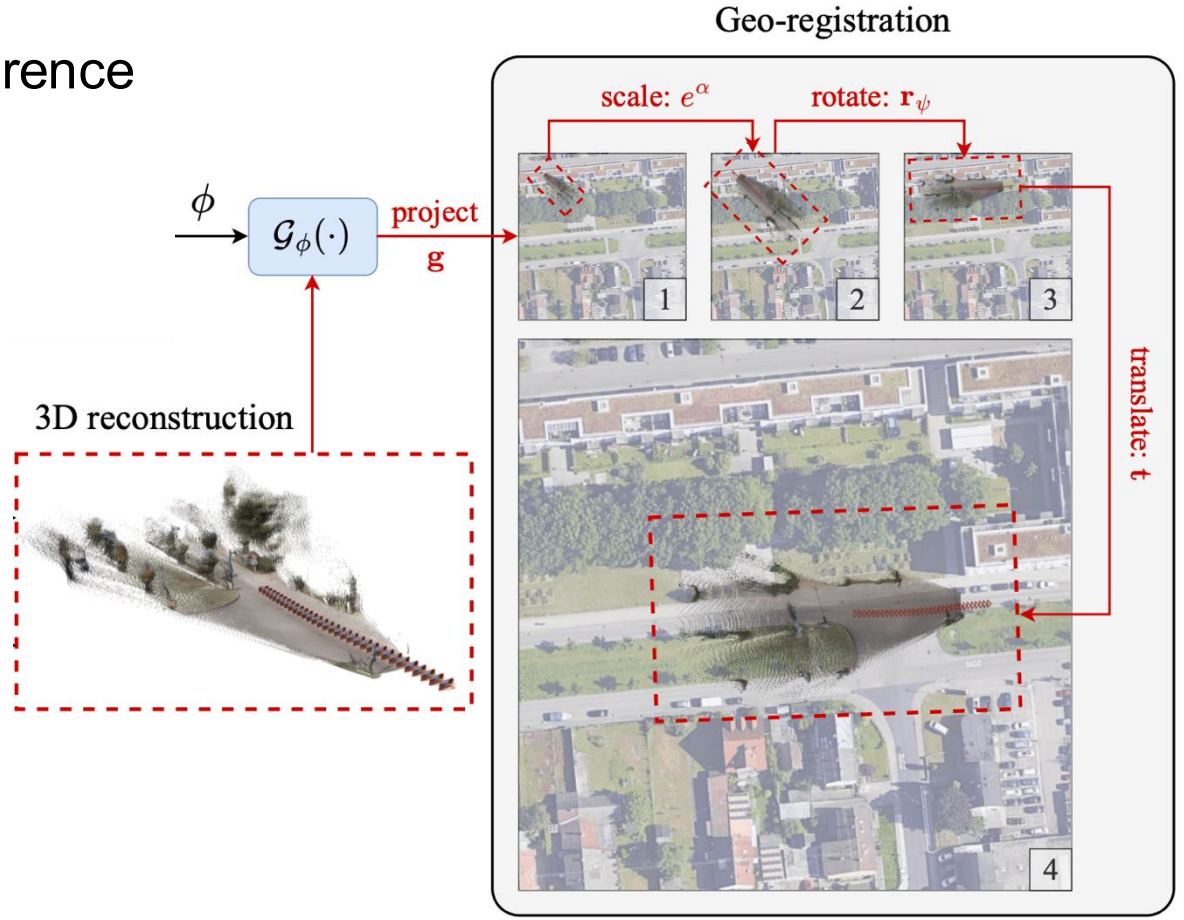


Architecture: **Geo-Registration**

- Map ground 3D reconstruction into satellite reference

1. Project 3D points into ground plane with $\mathbf{g}$
2. Scale to metric with $s = e^{\alpha}$
3. Rotate yaw using $\mathbf{r}_{\psi}$
4. Translate points using $\mathbf{t}$

- Each point is mapped into GPS coordinates



Georegistration Equations

Georegistration Algorithm

Let $\mathbf{e}_x = [1, 0, 0]^\top$ be the x-axis in the predicted ground reference frame

Define $\mathbf{a}$ as the projection of $\mathbf{e}_x$ into the ground plane

$$\tilde{\mathbf{a}} = \mathbf{e}_x - (\mathbf{e}_x^\top \mathbf{g})\mathbf{g}, \quad \mathbf{a} = \tilde{\mathbf{a}}/\|\tilde{\mathbf{a}}\|.$$

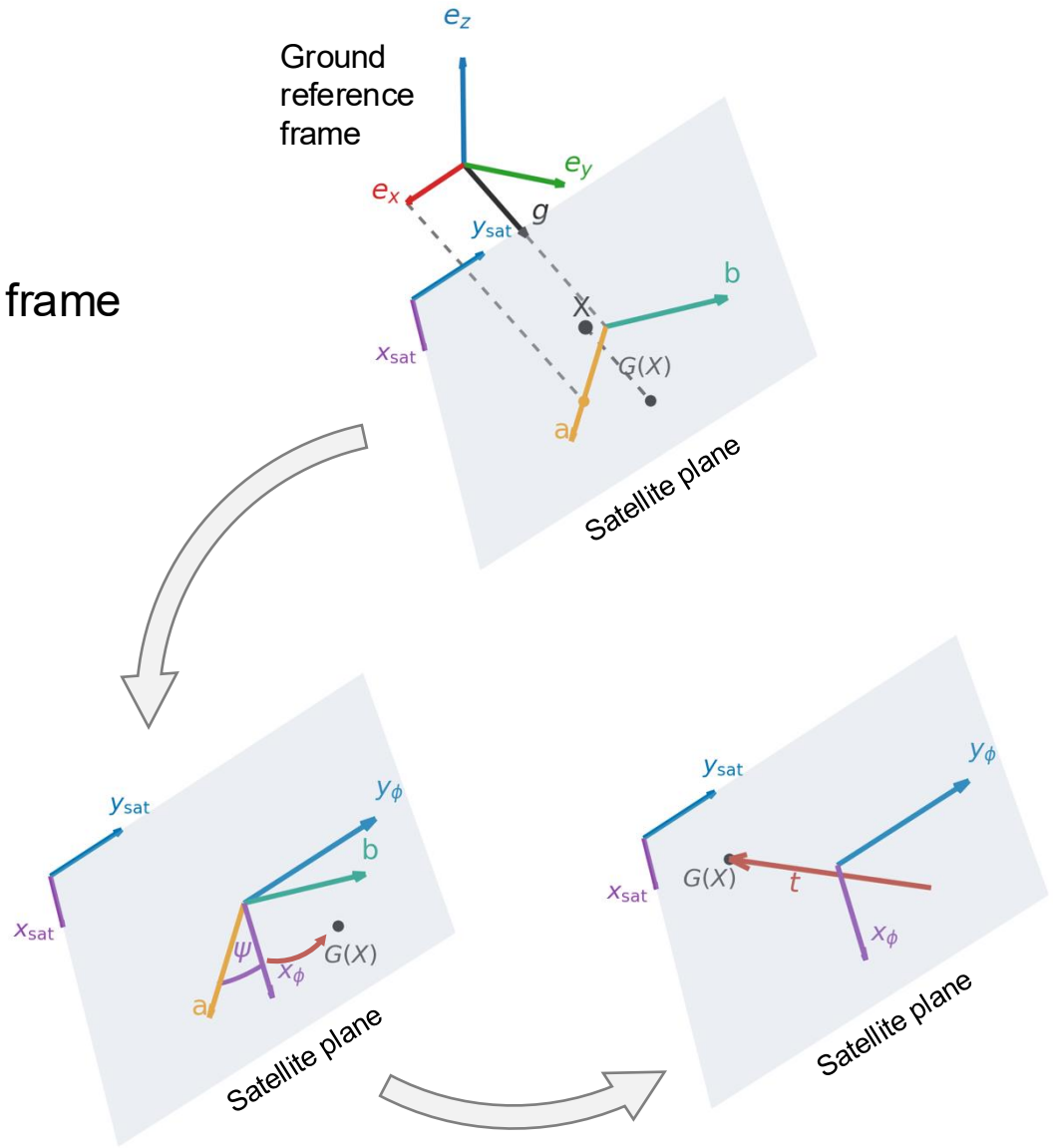
And $\mathbf{b}$ as the orthogonal vector: $\mathbf{b} = \mathbf{g} \times \mathbf{a}$.

Rotate the basis using the predicted yaw:

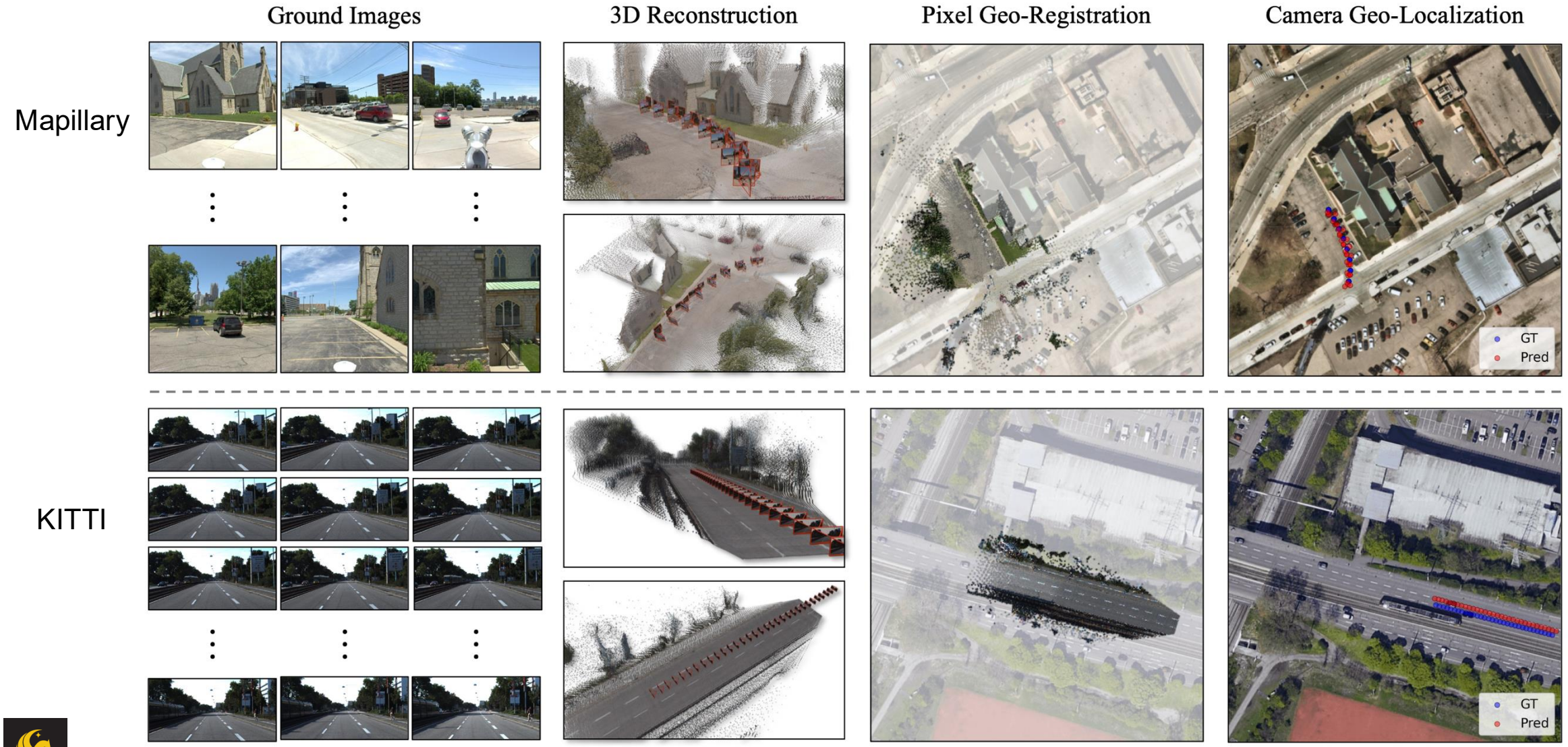
$$\mathbf{x}_\phi = \cos\psi\, \mathbf{a} + \sin\psi \mathbf{b}, \quad \mathbf{y}_\phi = -\sin\psi\, \mathbf{a} + \cos\psi\, \mathbf{b}.$$

Then, for any 3D point $\mathbf{X}$:

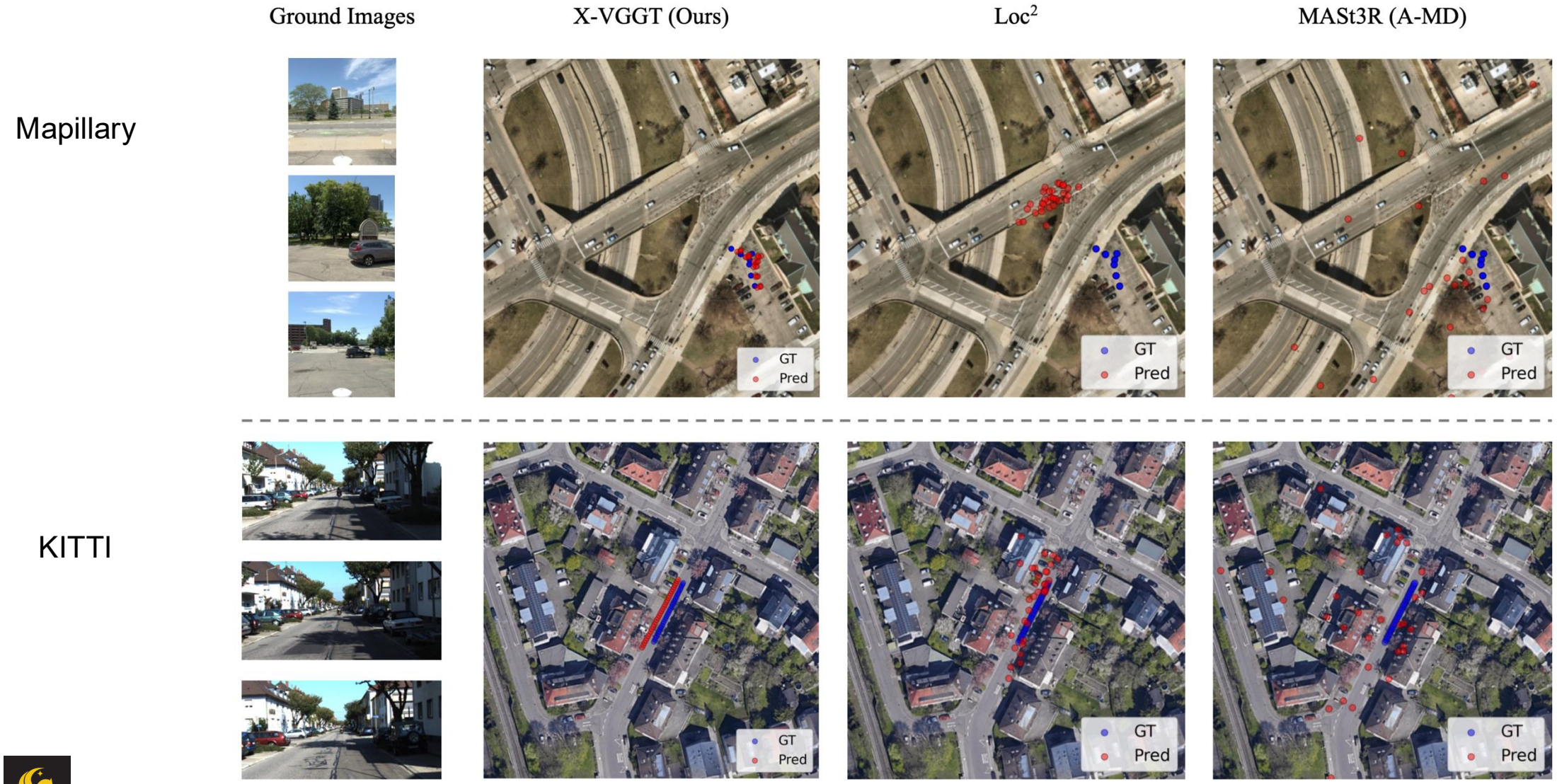
$$\mathcal{G}_\phi(\mathbf{X}) = s \begin{bmatrix} \mathbf{x}_\phi^\top \mathbf{X} \\ \mathbf{y}_\phi^\top \mathbf{X} \end{bmatrix} + \mathbf{t}.$$



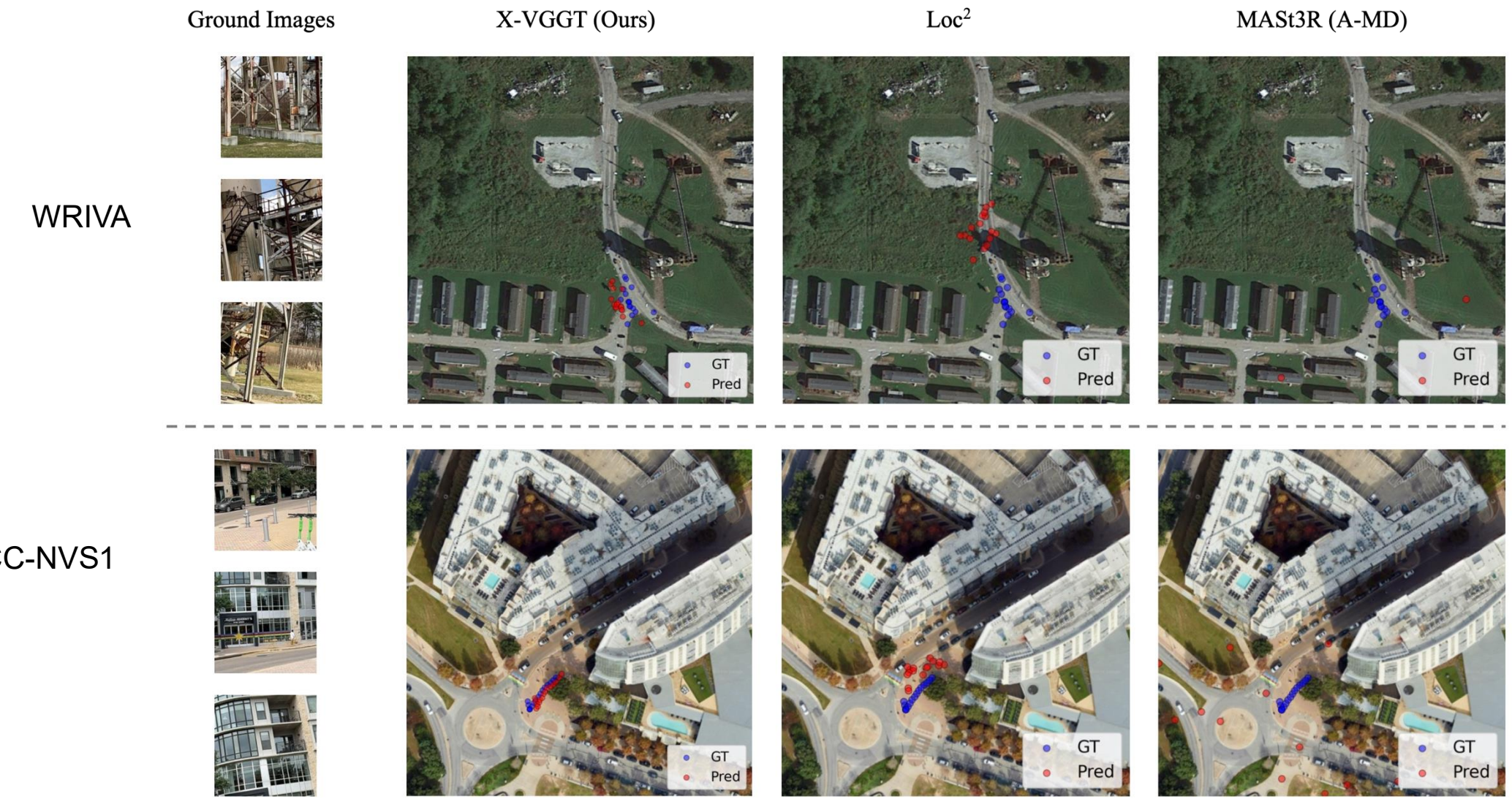
Qualitative Results



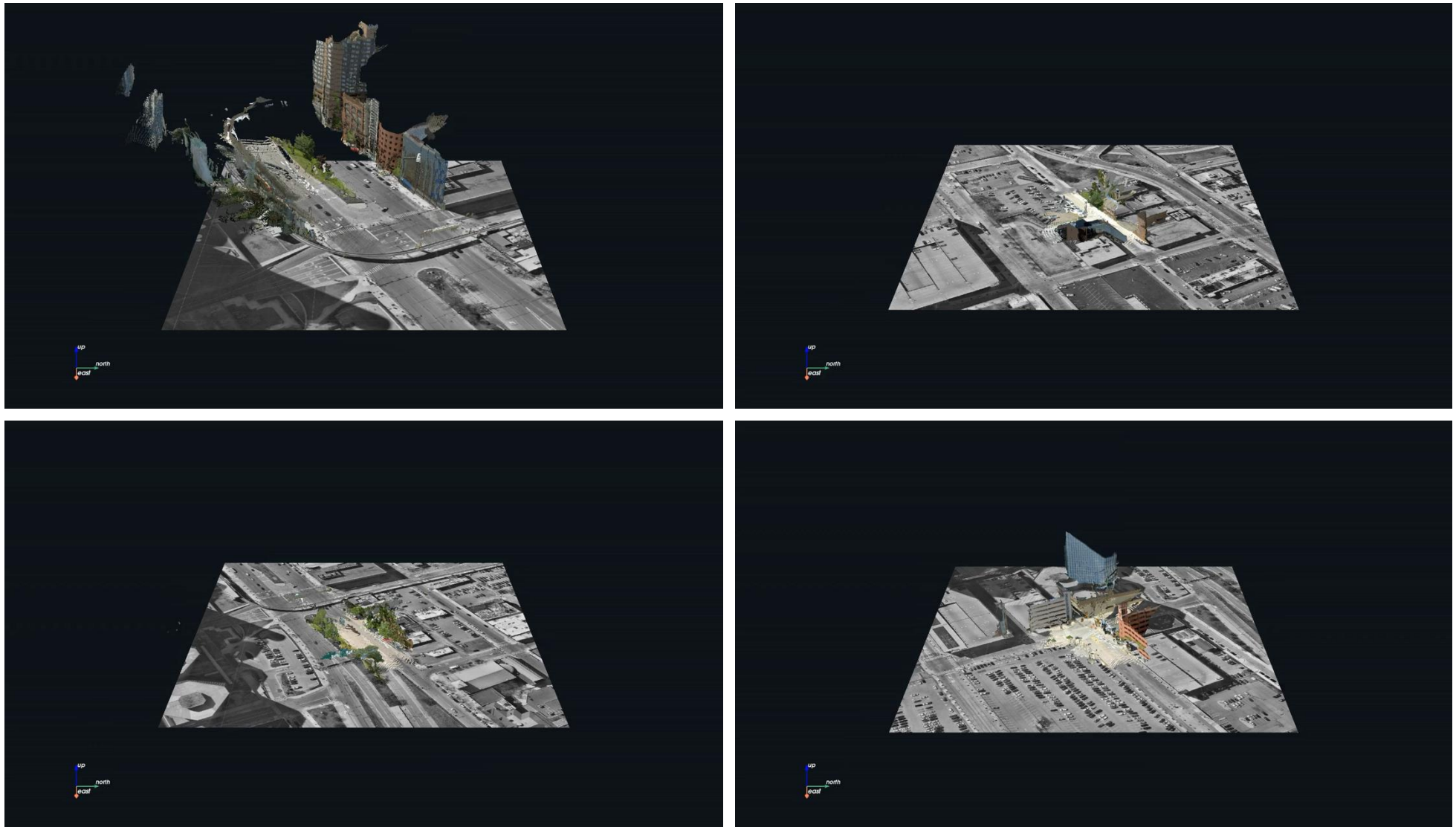
Qualitative Results



Qualitative Results



Qualitative Results



Quantitative Results: **Camera Geo-localization**

- Measured: location (Loc, meters) and orientation (Orien, deg) errors
- Lower is better

Method	KITTI		Metropolis		WRIVA		ACC-NVS1	
	Loc. (m)	Orien. (°)	Loc. (m)	Orien. (°)	Loc. (m)	Orien. (°)	Loc. (m)	Orien. (°)
FG^2 [55]	39.81	89.00	40.26	89.73	36.47	85.80	36.51	77.58
Loc^2 [56]	32.00	69.99	37.61	84.09	36.30	76.16	39.96	80.81
MASt3R [21]	1103.32	62.22	1028.05	79.83	1446.36	88.89	952.52	86.13
VGGT [46]	130.05	32.17	408.13	60.77	455.36	63.00	168.45	77.54
A-MD [43]	104.95	24.68	154.35	75.82	412.00	79.24	190.72	74.84
VGGT + Loc^2	31.46	81.16	37.06	89.19	35.16	87.41	37.07	89.41
X-VGGT (Ours)	**24.62**	**22.50**	**14.35**	**17.59**	**26.05**	**40.64**	**24.49**	**42.99**

Quantitative Results: **Georegistration** and **Relative Pose**

- Measured: pixel-wise metric error (m)
- Lower is better

Method	KITTI	Metropolis
Loc^2 [56]	43.96	88.96
MASt3R [21]	900.85	1500.67
VGGT [46]	146.35	640.16
A-MD [43]	62.07	319.15
VGGT + Loc^2	178.45	50.54
X-VGGT (Ours)	**25.52**	**17.84**

- Measured: local rotation and translation AUC
- Higher is better

Method	KITTI	Metropolis	WRIVA	ACC-NVS1
MASt3R [21]	6.64	5.05	0.03	6.25
VGGT [46]	83.29	**55.31**	16.90	67.96
A-MD [43]	21.58	9.20	4.13	8.55
X-VGGT (Ours)	**83.90**	51.34	**21.17**	**72.91**

UCF | Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

Georeferenced Cross-View 3D Reconstruction from Ground and Satellite Images

David G. Shatwell, Mubarak Shah

*Accepted in **NeurIPS 2026***