Towards Spatial Intelligence for Geo-locating Ground Images onto Satellite Imagery

WRIVINDER

C. Gudavalli · T. M. Mohammed · A. Yadav · A. V. Bhaskar · H. Prajapati · C. Peng · R. Chellappa · S. Chandrasekaran · B. S. Manjunath

Mayachitra, Inc. | *Johns Hopkins University*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026) · Denver, Colorado · June 3–7

arXiv: 2602.14929 · github.com/Mayachitra-Inc/wrivinder

arXiv GitHub

Acknowledgement: This research is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number 140D0423C0076.

Agenda

- Motivation and Problem Statement
- Main Contributions
- MC-Sat Dataset
- Methodology
  - Wrivinder System Overview
  - Deep Template Matcher
- Qualitative and Quantitative Results
- Summary / Conclusions

Motivation

🛰️ **Extreme Viewpoint Gap**: Ground photos and satellite maps show the same place from completely different altitudes, angles, and scales — making direct feature matching unreliable.

🧭 **Direct Matching is Unstable**: State-of-the-art matchers like SIFT, SuperPoint, and RoMA all fail when applied directly across this viewpoint gap.

🎯 **Geometry as the Bridge**: Use SfM + 3DGS to reconstruct a 3D scene and render a top-down zenith view — then align it to satellite imagery. No paired training data needed.

Motivation

🛰️ **Extreme Viewpoint Gap**: Ground photos and satellite maps show the same place from completely different altitudes, angles, and scales — making direct feature matching unreliable.

🧭 **Direct Matching is Unstable**: State-of-the-art matchers like SIFT, SuperPoint, and RoMA all fail when applied directly across this viewpoint gap.

🎯 **Geometry as the Bridge**: Use SfM + 3DGS to reconstruct a 3D scene and render a top-down zenith view — then align it to satellite imagery. No paired training data needed.

Ground-to-Satellite Alignment: Wrivinder reconstructs 3D geometry from ground images and aligns zenith renderings to satellite imagery.

Agenda

- Motivation and Problem Statement
- Main Contributions
- MC-Sat Dataset
- Methodology
  - Wrivinder System Overview
  - Deep Template Matcher
- Qualitative and Quantitative Results
- Summary / Conclusions

Main Contributions

🧱 MC-Sat Dataset: First benchmark linking multi-view ground imagery with geo-registered satellite tiles — 15 scenes, ~20K images, 4 continents.

🧱 Wrivinder Pipeline: Zero-shot framework — SfM + 3DGS + monocular depth + self-supervised DTM — recovers GPS coordinates with no ground-satellite training pairs.

🎯 Deep Template Matcher (DTM): Siamese ResNet-18 trained at test-time on synthetic satellite crop pairs — localizes the zenith render within the satellite tile with no ground-truth labels.

Ground-to-Satellite Alignment: Wrivinder reconstructs 3D geometry from ground images and aligns zenith renderings to satellite imagery.

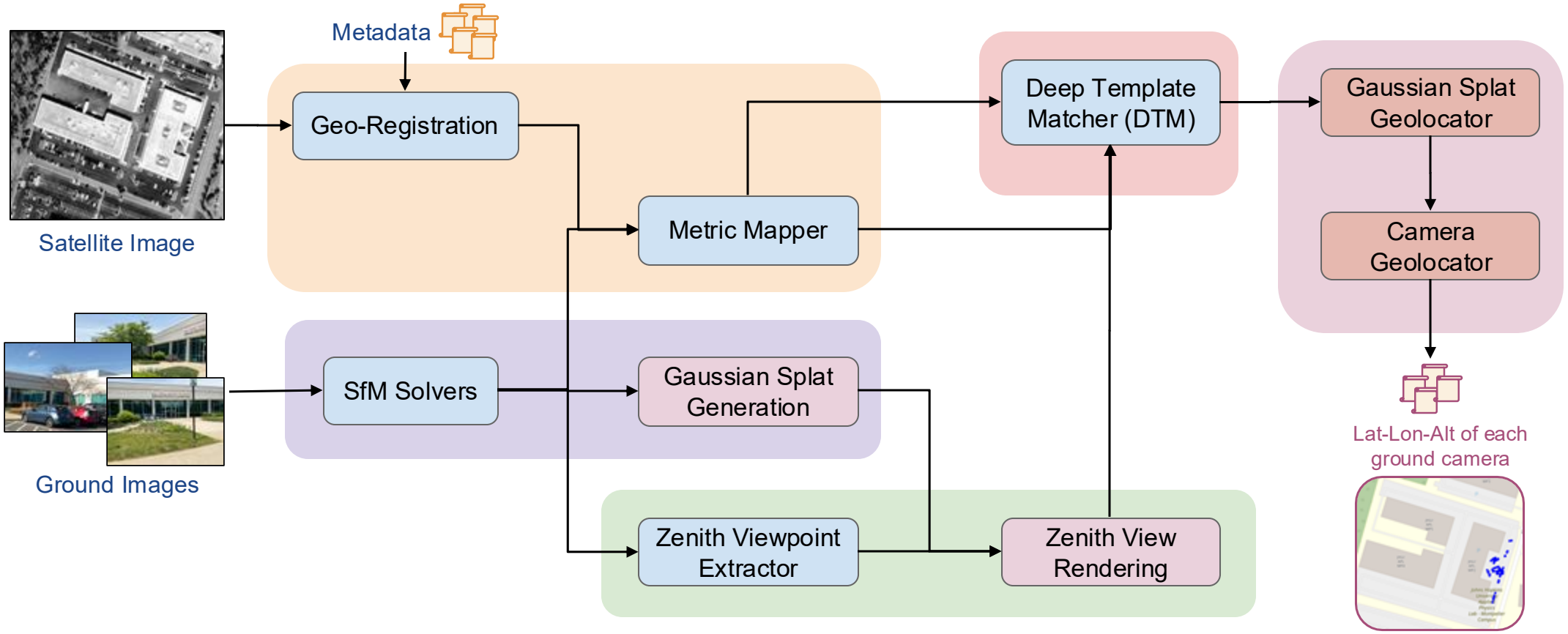
MC Sat Dataset

15	20K	4
Scenes	Images	Sources

Dataset	Scenes	Images	Type
ULTRAA	3	1K	Ground
VisymScenes	149	258K	Ground
ACC-NVS	6	148K	Gnd+Air
JHU-Ames	1	1.7K	Gnd+Air

Sample scenes from MC-Sat: satellite view (center) with surrounding ground images.

Methodology



Five-stage zero-shot pipeline: ground images and a satellite tile in → metrically accurate Lat-Lon-Alt for every camera out

Methodology

The pipeline generates a series of sophisticated intermediate representations to bridge the gap between street-level captures and top-down georeferencing.

(a) Ground Images
Unordered photos captured around a building from street level.

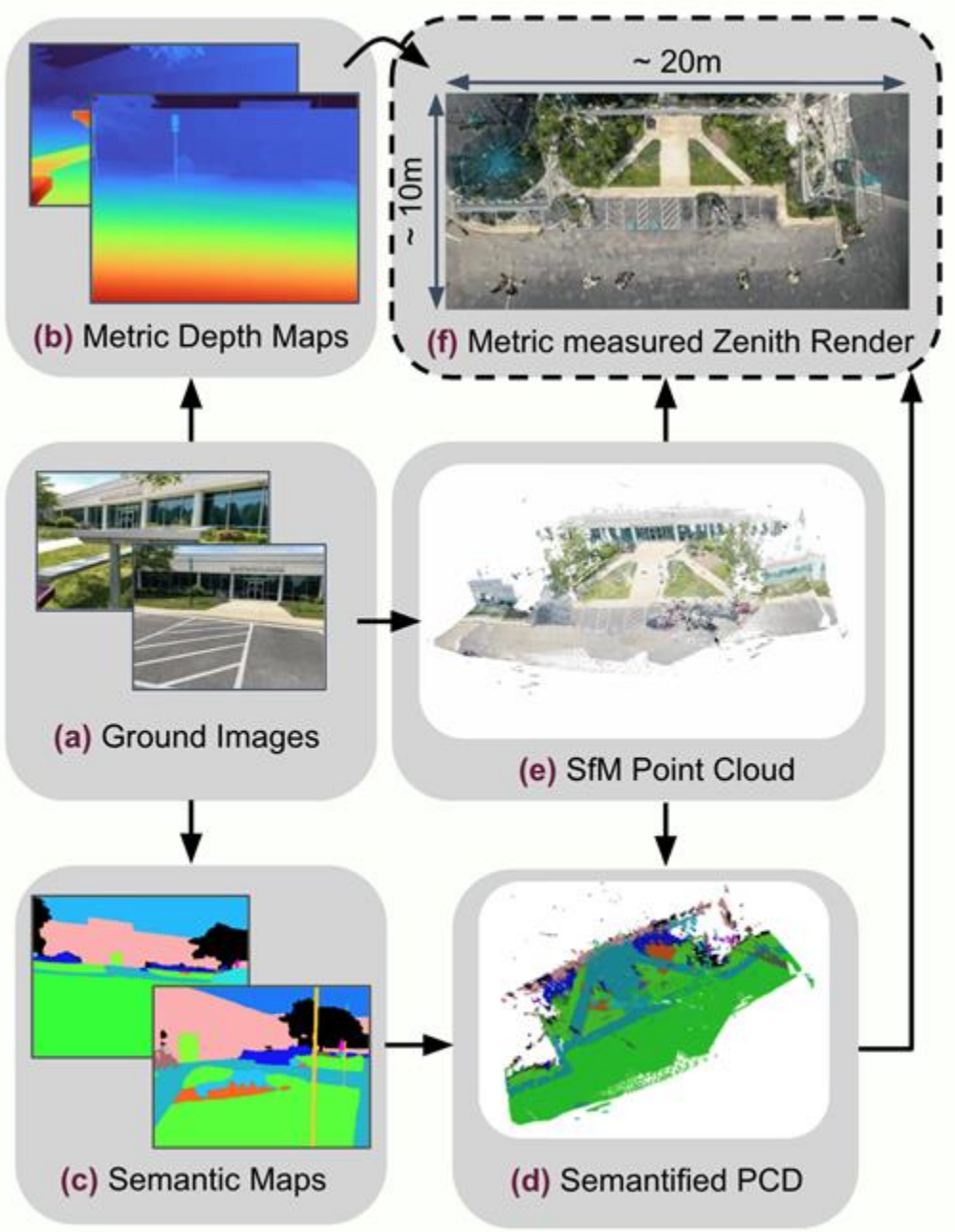
(b) Metric Depth Maps
Monocular depth predictions (DepthPro/PatchFusion) used to recover real-world scale.

(c) Semantic Maps
Mask2Former segmentation identifying road, grass, pavement, and ground-plane classes.

(d) Semantified Point Cloud
SfM sparse 3D reconstruction with semantic labels propagated to each 3D point.

(e) SfM Point Cloud
Sparse geometry estimated by HLOC+COLMAP across all input views.

(f) Metric Zenith Render
3DGS rendered from directly above, with physical footprint annotated (~10m × 20m).



Key Intermediate Outputs of Wrivinder: Semantic maps, SfM reconstruction, metric depth estimation, and the resulting

Agenda

- Motivation and Problem Statement
- Main Contributions
- MC-Sat Dataset
- Methodology
  - Wrivinder System Overview
  - Deep Template Matcher
- Qualitative and Quantitative Results
- Summary / Conclusions

Quantitative Results

Dense scenes achieve near-meter-level localization, while large reconstructed areas remain challenging due to incomplete rooftop visibility and sparse zenith coverage.

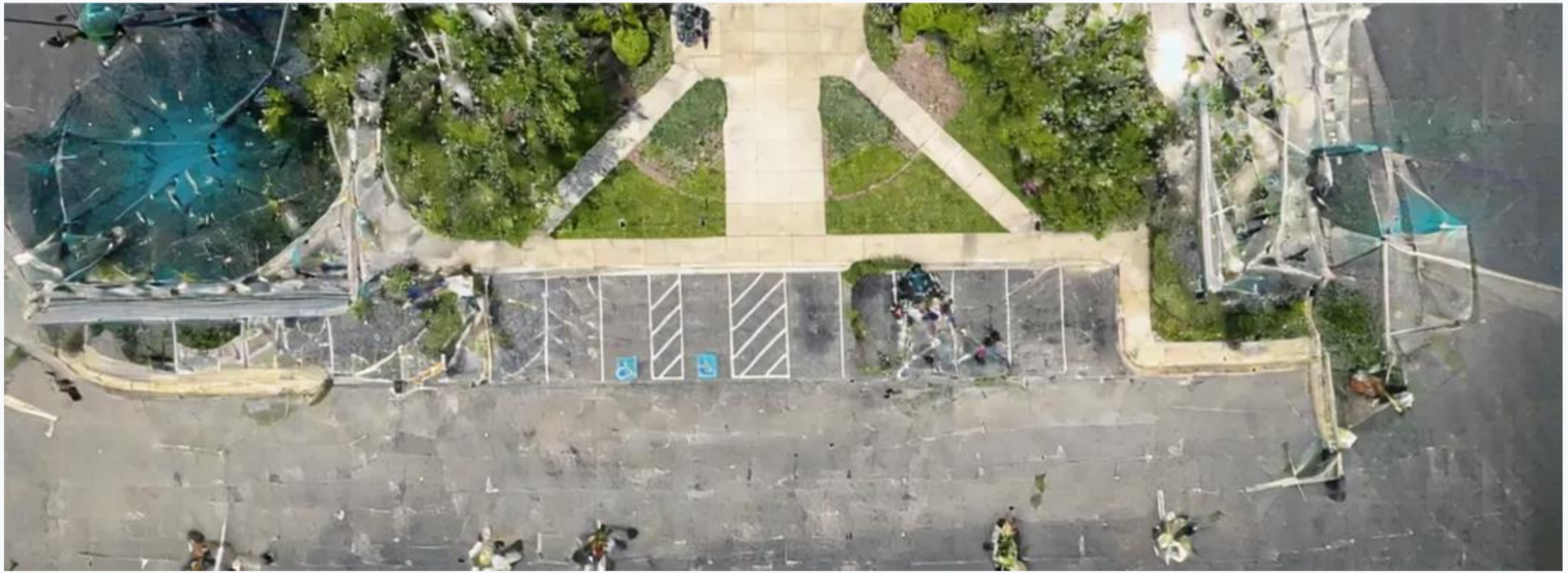
Scene	Type	Geo RMSE ↓
APL Front Door	Dense	1.96 m
APL Back Door	Dense	2.82 m
siteACC0003	Dense	3.02 m
MUTC A09	Large	18.86 m
MUTC A10	Large	17.82 m
siteSTR0003	Large	17.67 m
siteSTR0058 (US)	Large	11.88 m
siteSTR0059	Dense	32.13 m

< 6 m (near-meter) 6–20 m > 20 m

3DGS Zenith Render — APL Front Door

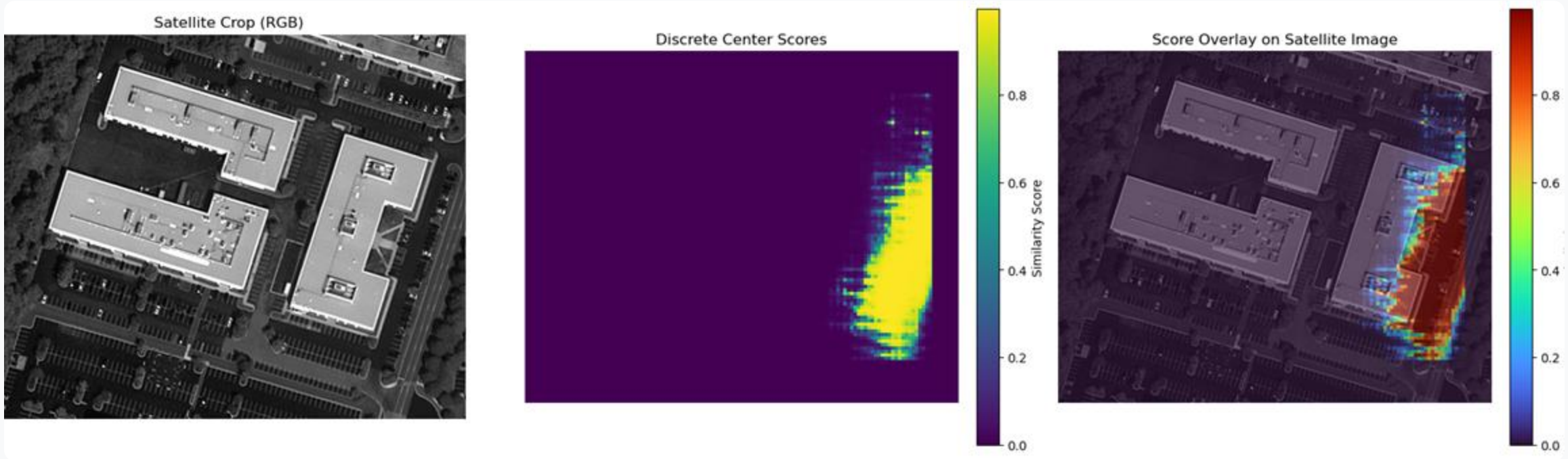
Standard 3DGS zenith render before Diffix3D inpainting. **Gaussian splat artifacts and floaters** are visible at the edges — these create noise in the zenith template used for satellite matching.

3DGS Zenith Render — After Diffix3D Post Processing



After Diffix3D inpainting, **splat artifacts are suppressed** and the zenith render shows clean structure: parking lot, courtyard, pathways, and landscaping — all reconstructed from ground-level photos only.

Deep Template Matcher — APL Front Door (s00)



01. Satellite Crop (RGB)

NAIP imagery at ~0.6 m/px resolution showing the target area from a top-down perspective.

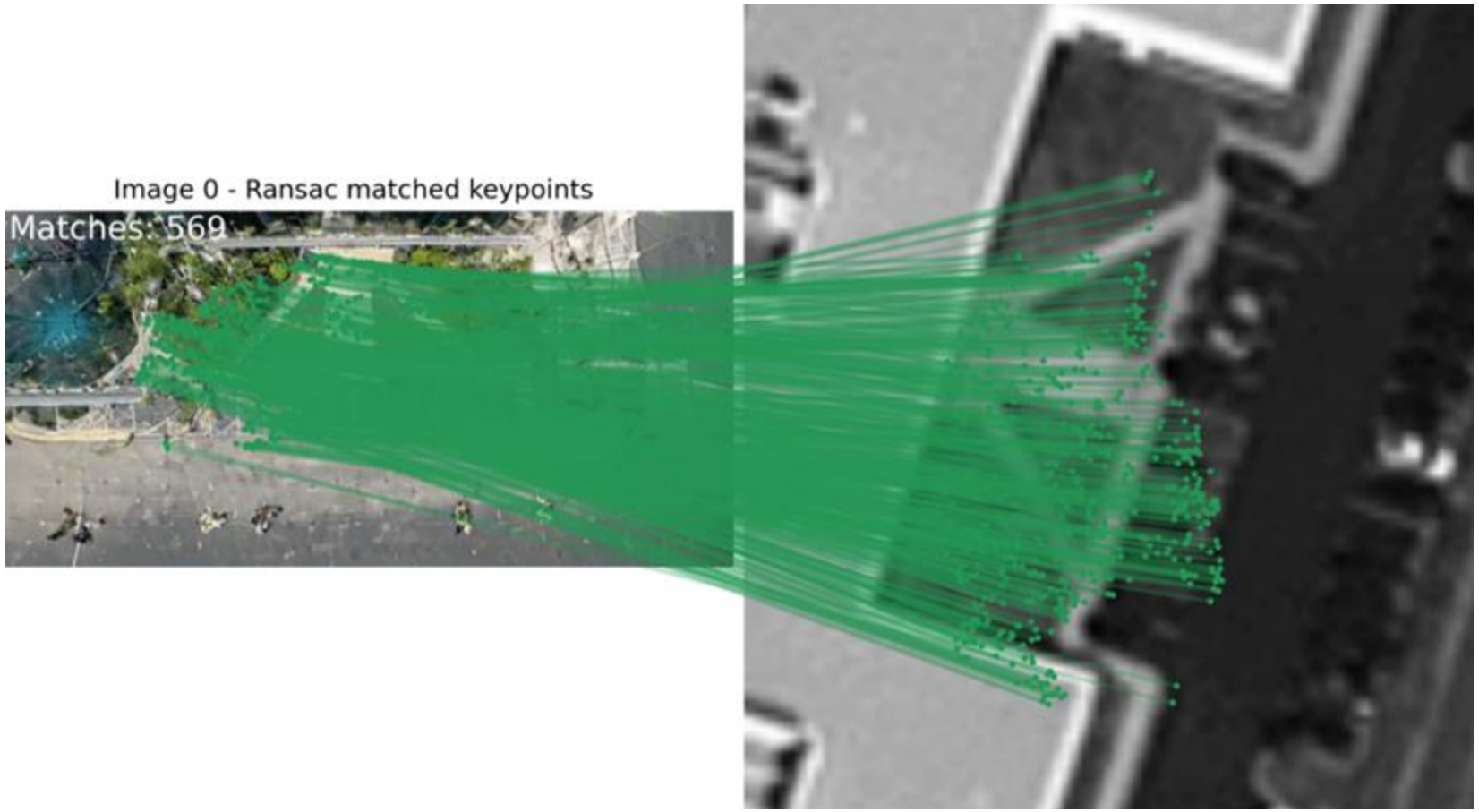
02. DTM Similarity Heatmap

Bright yellow cluster indicates peak response along the right edge, identifying the front entrance.

03. Score Overlay

Confirming localization over the correct region with a peak confidence score exceeding 0.8.

Gaussian Splat Geolocation: MatchAnything Model



APL Front Door

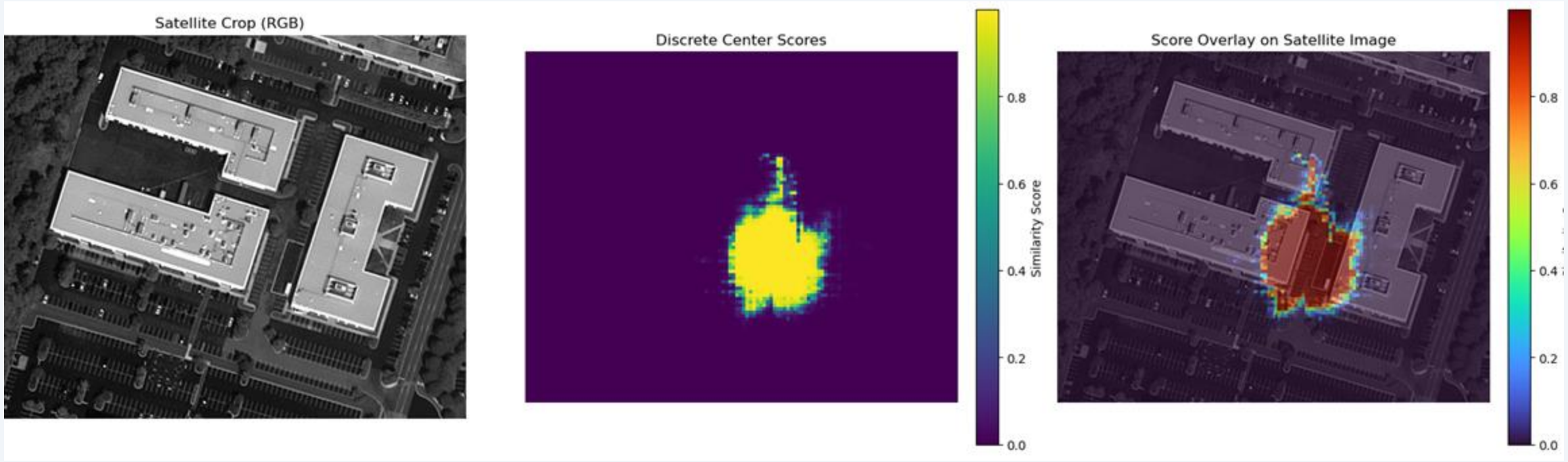
569 RANSAC-verified correspondences between the 3DGS zenith render (left) and the DTM-localized satellite crop (right).

Dense cross-modal matching via **MatchAnything-RoMA**.

These point pairs are back-projected through 3DGS → SfM to yield GPS coordinates.

Scene: APL Front Door — Performance of cross view point matcher on 3DGS zenith render (left) and DTM-localized satellite crop (right).

Deep Template Matcher — APL Back Door (s01)



01. Satellite Crop (RGB)

Same satellite crop as Front Door scene — building visible from above.

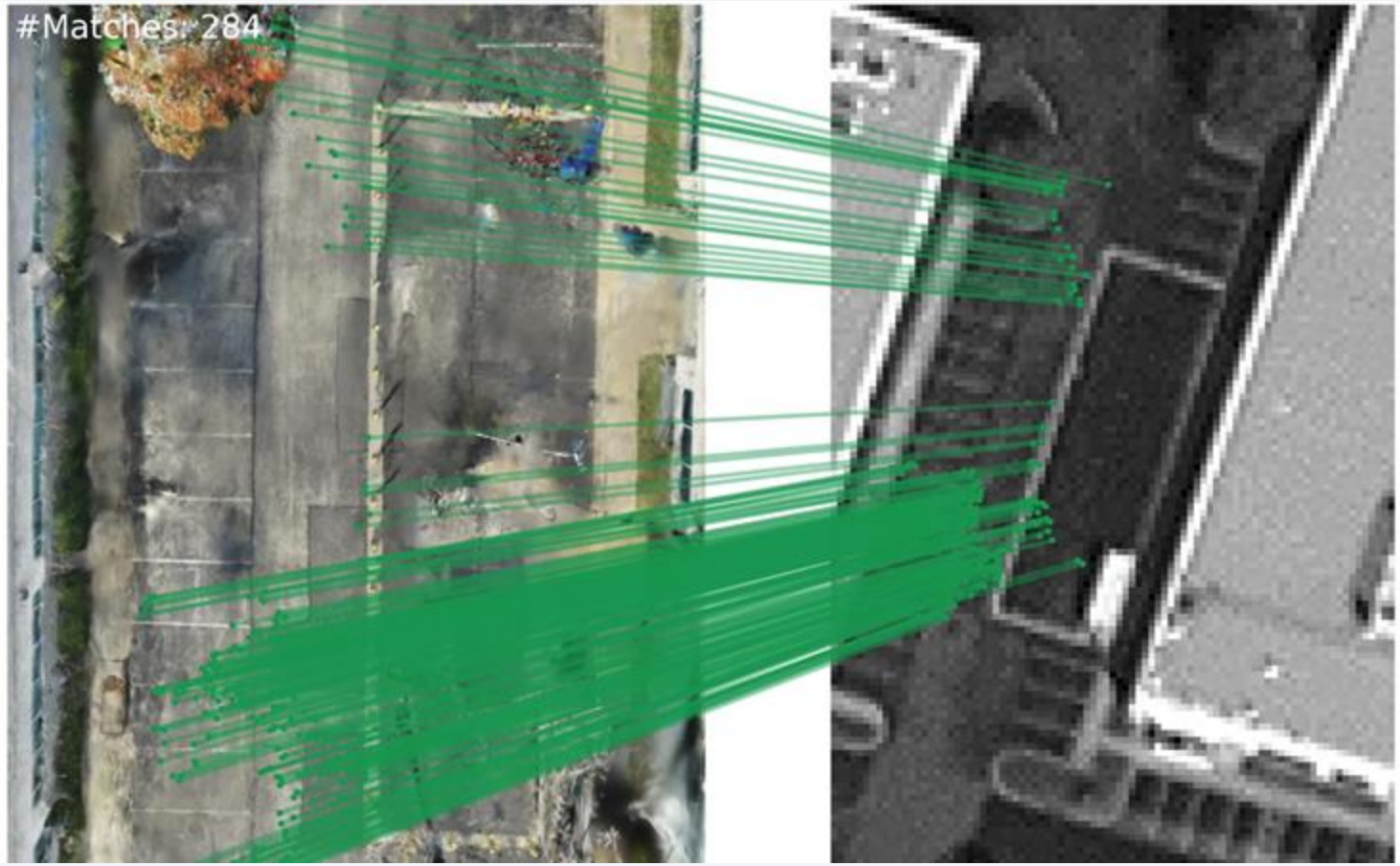
02. DTM Similarity Heatmap

DTM heatmap shows a tight, well-localised blob (high confidence) — the compact response indicates the Back Door cluster is spatially distinct from the Front Door result.

03. Score Overlay

Score overlay confirms localization over the rear parking area.

Gaussian Splat Geolocation: MatchAnything Model



Visualization of correspondences between ground views and satellite imagery for geolocation.

APL Back Door

284 RANSAC-verified matches across two ground-camera views.

Dense cross-modal matching via **MatchAnything-RoMA**.

Despite a different capture perspective from the building rear, the model recovers dense, geometrically consistent correspondences.

Result: 2.82 m mean geolocation error

Agenda

- Motivation and Problem Statement
- Main Contributions
- MC-Sat Dataset
- Methodology
  - Wrivinder System Overview
  - Deep Template Matcher
- Qualitative and Quantitative Results
- Summary / Conclusions

Summary

🌍 Zero-shot ground-to-satellite alignment — no paired training data, no fine-tuning, no GPS required at inference

📐 SfM + 3DGS bridges the viewpoint gap — photorealistic zenith renders match directly against satellite imagery

📍 Near-meter accuracy on dense scenes (APL Front Door: 1.96 m, Back Door: 2.82 m) — sub-20 m on large-area reconstructions

GITHUB

A RELATED OPEN CHALLENGE (WRIVA CVGL)