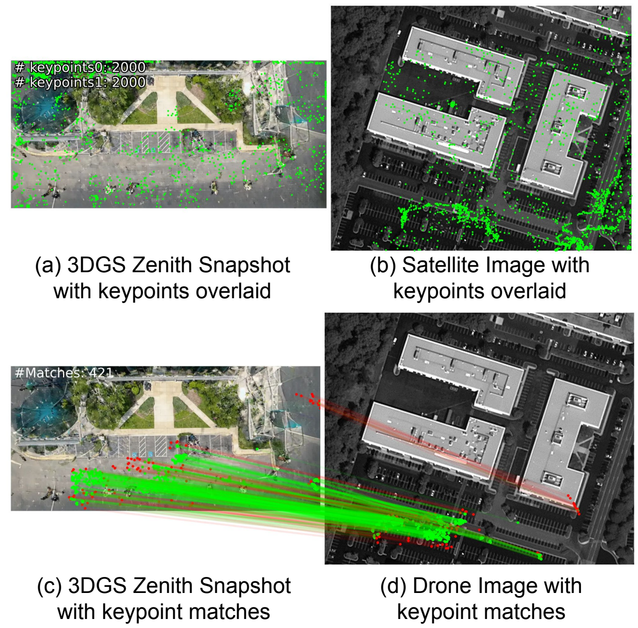
`wrivinder`: Towards Spatial Intelligence for Geo-locating Ground Images onto Satellite Imagery

Supplementary Material

1. Overview

This supplementary material provides additional analyses, qualitative results, and supporting details that complement the main paper. We first examine the challenges of aligning 3DGS zenith renders to satellite imagery, including the limitations of traditional keypoint matchers and several self-supervised attempts based on DINOv3 features (Sec. 2). We then describe the construction of the synthetic training dataset used for our Siamese ResNet matcher, highlighting the role of metric-aware satellite cropping and blobby-jitter augmentations (Sec. 2.3). Next, we present qualitative correspondence results obtained with the MatchAnything module once coarse localization is established by the DTM stage (Sec. 3). Finally, in Section 4 we provide an expanded background discussion of prior work related to ground-to-satellite geolocation, offering additional context that may help readers better understand the broader research landscape surrounding this problem.



(a) 3DGS Zenith Snapshot with keypoints overlaid
(b) Satellite Image with keypoints overlaid
(c) 3DGS Zenith Snapshot with keypoint matches
(d) Drone Image with keypoint matches

Figure 1. Direct feature matching between **(a)** 3DGS zenith render and **(b)** satellite image using RoMA. Although a few keypoints are detected(**c, d**), the majority of matches are unreliable or geometrically inconsistent, underscoring the difficulty of applying off-the-shelf matchers to this cross-view, cross-modal setting.

2. Zenith Render to Satellite Image Matching

2.1. Performance with Point Matchers

Traditional sparse feature matchers—such as SIFT, SuperPoint, LoFTR, or RoMA—are designed for viewpoint-consistent or modality-similar image pairs. When applied directly to the 3DGS zenith render and the corresponding satellite image, these matchers fail to produce stable or geometrically meaningful correspondences. As shown in Fig. 1, keypoints detected on the zenith snapshot often lie on splat boundaries, blurred regions, or view-dependent artifacts introduced during 3DGS rendering. Conversely, satellite imagery contains rooftops, tree canopies, shadows, and high-altitude structures that are either absent or heavily distorted in the zenith render.

As a result, naïve point matching yields sparse, inconsistent, and frequently incorrect correspondences (Figs. 1c–d). These mismatches arise from the severe cross-modal and cross-viewpoint gap between the inputs, and they ultimately prevent reliable geometric alignment. This motivates the need for a dedicated alignment mechanism and justifies the design of our test-time Deep Template Matcher (DTM), which operates at the level of dense appearance similarity rather than sparse keypoints.

2.2. Several Explored TTT/SSL Architectures for Deep Template Matching

We began by evaluating whether DINOv3 features could provide a useful signal for localizing the zenith render within a satellite tile. As shown in Fig. 3, DINOv3 produces strong semantic activations: a forest patch consistently highlights all forest regions, and a road patch activates along every road segment. Given that DINOv3 embeddings seem to be generalizable to satellite images as well, we explored several self-supervised (SSL) and test-time training (TTT) variants built on DINOv3 embeddings, aiming to learn a crop–crop similarity function without requiring ground-truth zenith coordinates. Figure 2 summarizes these attempts. Although Experiments 3 and 4 introduce multi-pass and noise-perturbed token extraction schemes, their heatmaps (Fig. 2c–d) fail to produce a reliable localization peak. In contrast, the Siamese ResNet model adopted in the main paper (Fig. 2e) yields a clean and well-defined match, highlighting the limitations of the DINOv3-based SSL formulations.

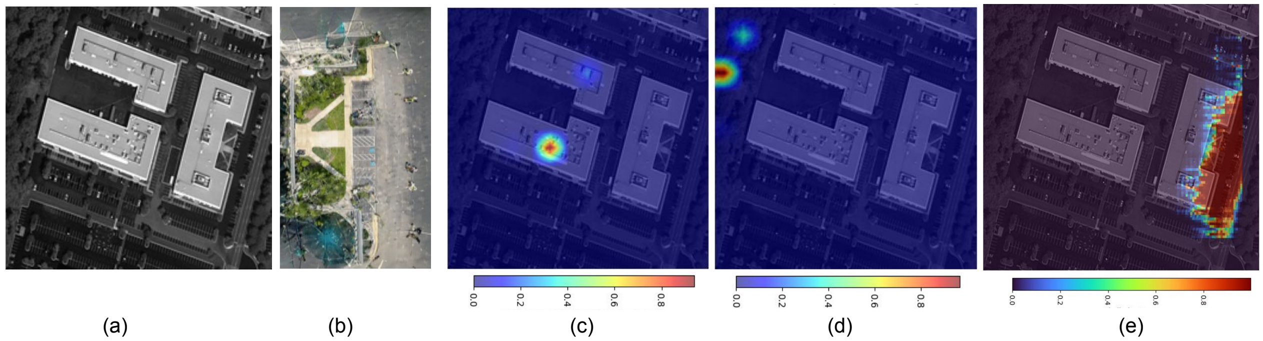


Figure 2. Comparison of our DTM experiments. The heatmaps in (c) and (d) correspond to Experiment 3 and Experiment 4, illustrating that the DINOv3-based SSL approaches fail to produce accurate localization peak. In contrast, the Siamese ResNet model used in the main paper (e) yields a clean and well-defined match, demonstrating the effectiveness of the final architecture.

Context. Our objective is to perform self-supervised learning (SSL) for identifying the *zenith render* within a satellite image. Since ground-truth coordinates of the zenith-render location are unavailable during training, we treat one crop in each sampled crop-pair as a *proxy* zenith view and supervise the relationship between this crop and another via IoU-based labels. Despite exploring several architectural directions, none of the variants produced a stable zenith-localization signal.

Experiment 1: Single Forward Pass + Binary Labels. Patch (token) embeddings are extracted from a single DINOv3 forward pass over the full satellite image. Multiple crops of varying scale and aspect ratio serve as training anchors, and for each sampled pair (i, j) we treat one crop as a proxy zenith-render and the second as a comparison crop. Tokens for both crops are obtained by mapping crop coordinates to the full-image token grid. We assign:

$$\text{Positive if IoU} > 0.3, \qquad \text{Negative if IoU} < 0.1,$$

with binary labels $y \in \{0, 1\}$. A lightweight Transformer layer cross-attends the query-crop tokens to the partner-crop tokens (key/value). The aggregated representation is passed to an MLP. The model is trained with binary cross-entropy:

$$\mathcal{L} = \text{BCE}(\hat{y}, y).$$

Experiment 2: Single Forward Pass + Soft IoU Labels. Identical to Experiment 1, but instead of binary supervision, the pair label is the continuous IoU. The cross-attention module and MLP remain unchanged. Optimization uses L2 regression:

$$\mathcal{L} = \|\hat{y} - y\|_2^2.$$

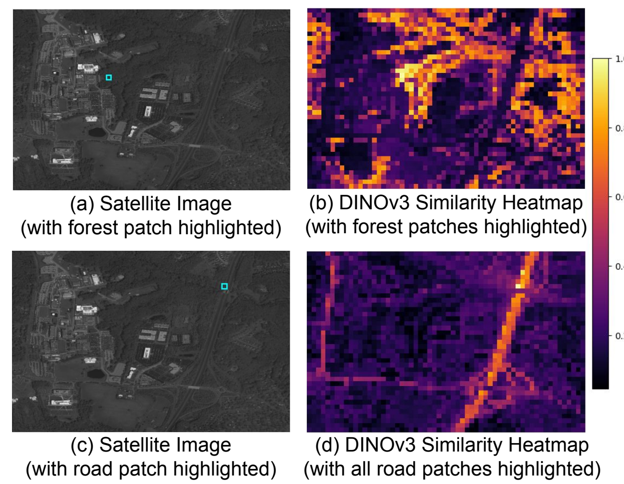


Figure 3. Similarity heatmaps produced by DINOv3 for forest and road patches. In both cases, the highlighted patch (a, c) triggers widespread activations across all semantically similar areas in the image (b, d).

Experiment 3: Multiple Forward Passes + Soft IoU Labels. To reduce misalignment introduced by token-mapping from crops to the full-image token grid, we use an asymmetric extraction scheme. For each pair, one proxy crop uses mapped tokens from the full-image forward pass, while the other crop is forwarded independently through DINOv3 to obtain context-specific tokens. Labels again correspond to the IoU of the two crops. A cross-attention Transformer and MLP predict the IoU via:

$$\mathcal{L} = \|\hat{y} - y\|_2^2.$$

Experiment 4: Multiple Forward Passes + Noisy Crops + Soft IoU Labels. This variant extends Experiment 3 by adding noise to the second crop before forwarding it through DINOv3. The aim was to explore whether noise-

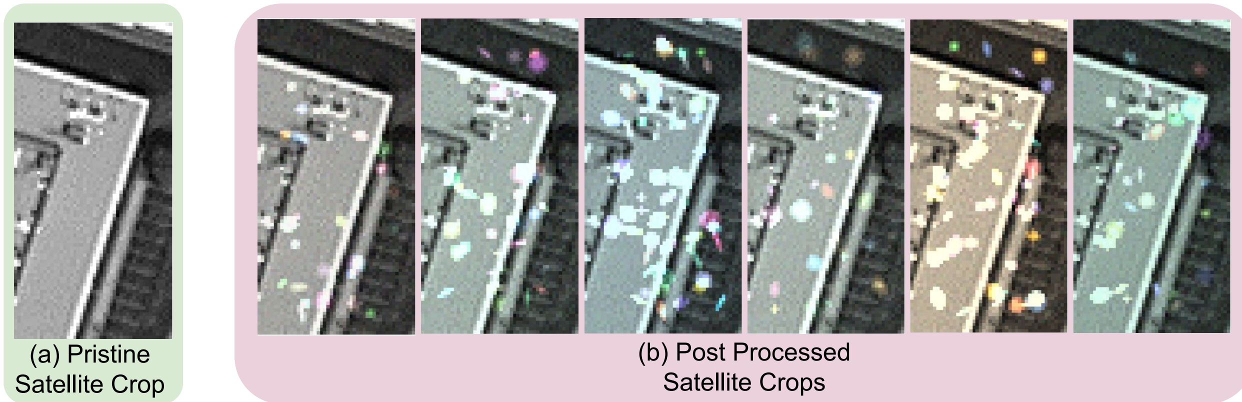


Figure 4. A clean satellite patch (a) and its corresponding blobby-jittered variants (b), used to train the Siamese ResNet matcher.

based perturbation (TTT-like) could induce more stable, invariant representations that help identify a robust proxy zenith signal. The rest of the pipeline remains unchanged, using IoU regression:

$$\mathcal{L} = \|\hat{y} - y\|_2^2.$$

Potential Reasons for Failure. Despite several architectural modifications, the self-supervised framing did not provide a usable zenith-localization signal. We attribute the failures to the following issues:

- **Weak proxy supervision.** Treating one crop as a proxy zenith render does not guarantee semantic or geometric correspondence to an actual zenith viewport. IoU between arbitrary crops does not reflect the latent notion of "zenithness" required for the task.
- **Misalignment of token representations.** Mapping crop coordinates to patch tokens (Experiments 1–2) introduces spatial quantization and positional inconsistency. Even in Experiments 3–4, independently computed tokens are not semantically aligned with those extracted from the full image.
- **Representation mismatch between crops.** DINOv3 features differ significantly depending on global context; isolated-crop tokens and full-image tokens come from different context windows, hindering stable cross-attention.
- **Insufficient signal in pairwise IoU.** IoU provides a purely geometric relation between crops; it does not correlate strongly with the appearance-based factors that define a zenith viewpoint.
- **Limited model capacity.** A small Transformer head may be insufficient to bridge the gap between DINOv3 patch semantics and the fine-grained geometric cues needed for zenith detection.
- **Noisy-crop instability.** Noise perturbations (Experiment 4) further distort the already fragile crop-based token representations, providing no useful TTT benefit.

2.3. Training Dataset for Siamese-ResNet Model

To train the Siamese ResNet matcher in a fully self-supervised manner, we construct synthetic patch pairs that mimic the visual characteristics of the 3DGS zenith render while still being derived entirely from the satellite image. The motivation is straightforward: zenith snapshots from 3DGS often contain view-dependent blur, splat boundaries, and localized "blobby" artifacts, whereas satellite imagery is significantly sharper and more photometrically consistent. The training set therefore needs to introduce controlled distortions that approximate these artifacts while preserving the underlying structure of the region.

A crucial component enabling this process is the Metric Mapper (Sec. 4.3 of the main paper). By estimating the approximate physical footprint of the reconstructed scene, the Metric Mapper provides a consistent crop size in pixel space that reflects the true metric scale of the zenith render. This metric prior greatly stabilizes the learning signal—when crops are sampled at the correct scale, the Siamese network sees patches that correspond to realistic 3DGS viewpoints, whereas incorrect scaling leads to degenerate or ambiguous supervision.

Figure 4 shows a clean satellite patch and several of its blobby-jittered variants. For each clean crop, we generate multiple perturbed versions by applying mild Gaussian blur, small brightness and contrast shifts, patchwise dropout, and scattered intensity blobs. These augmentations approximate the distortions commonly observed in 3DGS outputs and encourage the network to focus on regional layout and coarse geometry, rather than raw texture.

During training, the Siamese model receives the clean patch on one branch and a randomly chosen jittered version on the other. This simple strategy proved far more reliable than the DINOv3-based SSL formulations explored earlier, producing sharper similarity responses and more stable localization at test time.

3. Qualitative Results with MatchAnything Module for Gaussian Splat Geolocation

To refine the coarse localization produced by the DTM module, we apply the MatchAnything-RoMA matcher to the 3DGS zenith render and the corresponding satellite crop. While direct cross-modal matching fails at full-image scale (as shown earlier), constraining the search to the DTM-predicted region allows MatchAnything to recover dense and geometrically meaningful correspondences.

Figure 5 shows two examples from the APL Front and APL Backdoor scenes. In both cases, once the correct satellite window is identified, the matcher produces a large number of consistent point correspondences between the zenith snapshot and the satellite crop. These qualitative results confirm that, within a spatially localized region, MatchAnything can effectively serve as the fine-alignment stage of our Gaussian Splat geolocation pipeline.

4. Supplemental: Extended Related Work

This supplemental section provides additional context and detailed comparisons for the Related Work (Section 2) presented in the main paper.

Dataset Comparison and Availability

Existing CVGL Datasets. The landscape of cross-view geo-localization datasets reveals significant limitations for zero-shot evaluation. CVUSA [9] contains 35,532 training pairs and 8,884 test pairs from US cities, with all ground images being 360° panoramas. CVACT [6] extends this paradigm with 128,334 pairs from Canberra, Australia, maintaining the panoramic format but increasing test set diversity. VIGOR introduces same-area and cross-area evaluation protocols with limited field-of-view images from dash cameras. The recently proposed SetVL-480K [12] dataset represents the largest effort to date, containing 480,000 ground images from six global cities paired with satellite imagery.

However, SetVL-480K shares a critical limitation with all existing CVGL datasets: it is designed for supervised training with train-test splits from similar geographic distributions. Moreover, the SetVL-480K dataset and FlexGeo codebase are not publicly available at the time of writing, limiting reproducibility and community benchmarking. Existing datasets also predominantly feature one-to-few ground images per satellite crop (typically fewer than 10), whereas realistic SfM reconstruction scenarios benefit from denser image capture.

Our MC-Sat Dataset. In contrast, MC-Sat is explicitly designed for zero-shot evaluation of geometry-based localization methods. Our dataset provides:

- Multi-view ground imagery captured from diverse perspectives (average of 40+ images per location)
- Complete Structure-from-Motion reconstructions with camera poses
- Metric depth maps for each ground image
- Semantic segmentation masks with ground-plane annotations
- Geo-registered satellite imagery with known coordinate systems
- Ground-truth GPS alignments for quantitative evaluation

MC-Sat enables evaluation of methods that leverage 3D scene geometry rather than learned 2D feature matching. We will release the dataset, evaluation protocols, and baseline implementations to facilitate future research in geometric cross-view localization.

Detailed Comparison with Point Cloud Methods

Classical Point Cloud Alignment. Kaminsky et al.'s 2009 work [3] established the foundational approach for aligning SfM point clouds to overhead imagery. Their method introduces two complementary cost functions:

Edge Cost: Measures the proximity of projected 3D points to edges detected in the overhead image. This assumes that 3D structure boundaries will correspond to visible edges in the satellite view.

Free-Space Cost: Enforces that the space between the camera and visible 3D points must be free of occlusions. This visibility constraint helps reject incorrect alignments where points would be occluded by structures.

The optimization proceeds through coarse-to-fine search over the similarity transform space (rotation, translation, scale). For landmark buildings with strong geometric features, accuracy below 2 meters can be achieved when refined with Iterative Closest Point (ICP) and GPS priors.

Limitations of Sparse Point Clouds. While geometrically principled, sparse point cloud methods face fundamental limitations:

1. **Lack of Surface Continuity**: SfM produces discrete 3D points rather than continuous surfaces, making it difficult to establish dense correspondences with satellite imagery.
2. **Limited Appearance Information**: Point clouds carry minimal color or texture information, restricting matching to geometric features like edges and silhouettes.
3. **Unreliable Normals**: Point normals estimated from sparse neighborhoods are often noisy, especially in textureless regions or near depth discontinuities.
4. **Difficulty with Complex Structures**: Urban scenes with multiple building façades, vegetation, and ground clutter produce cluttered point clouds that are challenging to project cleanly to a 2D overhead view.

Wrivinder's Advancement. Our approach retains the geometric reconstruction foundation of Kaminsky et al. but

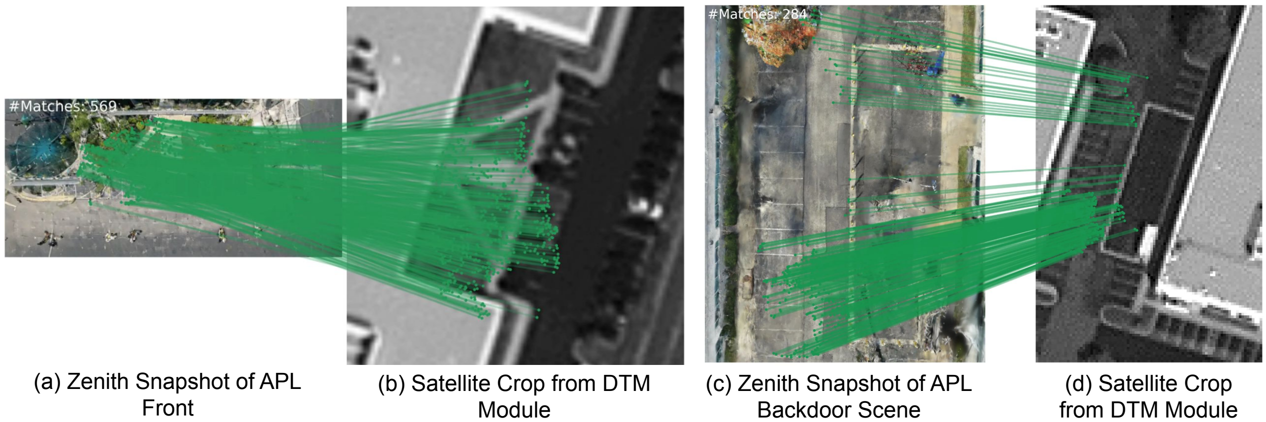


Figure 5. Example point correspondences produced by the MatchAnything module after coarse localization from the DTM stage. For both the APL front (a–b) and APL backdoor (c–d) scenes, the matcher recovers dense, geometrically meaningful matches between the 3DGS zenith render and the satellite crop.

addresses these limitations through 3D Gaussian Splatting:

- **Photorealistic Rendering**: 3DGS represents the scene as a collection of 3D Gaussians with learned appearance parameters, enabling high-quality novel view synthesis with full color and texture.
- **Continuous Representation**: The Gaussian representation provides implicit surface continuity, producing smooth renderings even from sparse input views.
- **Fast Optimization**: 3DGS trains 10-100× faster than NeRF (30 min - 2 hrs vs. 8-24 hrs), making it practical for deployment.
- **Real-Time Rendering**: At 90-150+ FPS, 3DGS enables interactive refinement and multiple hypothesis testing during alignment.

By rendering photorealistic zenith views, we transform the sparse geometric matching problem into a dense appearance-based matching problem, where modern learned matchers (e.g., LoFTR, RoMa) excel.

Neural Rendering: NeRF vs. 3DGS for Geo-Localization

NeRF-Based Approaches. Bredvik et al.'s STR method [1] demonstrates that Neural Radiance Fields can effectively bridge the ground-to-satellite viewpoint gap. Their pipeline:

1. Uses DeDoDe + RoMa for Structure-from-Motion initialization
2. Trains nerfacto-huge with depth regularization (8-24 hours per scene)
3. Renders synthetic nadir (overhead) views from the NeRF model
4. Establishes correspondences with satellite orthophotos using RoMa matcher with cyclic consistency filtering
5. Achieves 0.83-6.03 meter localization accuracy

STR's NeRF approach validates the core insight that neural rendering can synthesize views suitable for cross-view matching. However, NeRF's implicit representation incurs significant computational costs and struggles with real-time applications.

Why 3D Gaussian Splatting? We adopt 3DGS for several critical advantages:

Speed: Training 3DGS on typical outdoor scenes takes 30 minutes to 2 hours vs. 8-24 hours for NeRF. This 10-100× speedup makes iterative refinement and rapid deployment feasible.

Rendering Performance: 3DGS renders at 90-150+ FPS vs. 1-5 FPS for NeRF, enabling real-time visualization and interactive alignment refinement.

View Synthesis Quality. Our setting requires synthesizing views with a substantial viewpoint change (ground-to-zenith). In this regime, rendering quality depends strongly on the geometric consistency of the underlying representation. Due to its explicit Gaussian parameterization and rasterization-based rendering, 3DGS often produces sharper and more stable renderings under large viewpoint transformations compared to NeRF-based methods, while avoiding many of the optimization and inference costs associated with implicit volumetric representations.

Geometric Accuracy: Recent work (GeomGS [5], SplatLoc [7], GSplatLoc [10]) demonstrates that 3DGS provides superior geometric accuracy for camera localization when integrated with geometric constraints. The explicit Gaussian representation allows direct manipulation and better preserves geometric relationships.

Scalability: 3DGS's explicit representation enables better handling of large-scale scenes and integration with LiDAR or other metric sensors when available.

Comparison Summary. Both Wrivinder and STR share the zero-shot, metadata-free philosophy and the insight that neural rendering bridges viewpoint gaps. However, Wrivinder's 3DGS-based approach offers substantially faster reconstruction (10-100×), real-time rendering (30-150× faster), and often higher quality zenith synthesis, making it significantly more practical for real-world deployment.

Learned BEV Transformations vs. Explicit Geometry

Bird's-Eye-View Methods. The Panorama-BEV Co-Retrieval Network [8] represents the state-of-the-art in learned view transformation. Their approach:

- Converts panoramic street views to Bird's-Eye-View using ground plane assumptions
- Employs a dual-branch architecture: one for panorama-to-satellite retrieval, one for BEV-to-satellite retrieval
- Introduces CVGlobal dataset with non-aligned orientations and cross-temporal tests
- Achieves strong results on CVUSA (R@1: 96.7%), CVACT (R@1: 90.8%), and especially VIGOR cross-area (R@1: 67.2%)

The key innovation is learning an implicit neural transformation from ground perspective to overhead view, which is then matched using standard retrieval architectures. This approach achieves excellent accuracy on standard benchmarks.

Fundamental Differences. However, learned BEV methods differ from Wrivinder in several critical ways:

Training Dependency: BEV transformations are learned through supervised training on paired ground-satellite datasets. The neural network implicitly learns the geometric transformation through backpropagation and contrastive losses. Deploying to new regions requires either fine-tuning or assuming the learned transformation generalizes.

Geometric Assumptions: Most BEV methods assume a planar ground surface to perform the perspective transformation. This assumption breaks down in hilly terrain, multi-level structures (e.g., overpasses, elevated walkways), or complex 3D scenes.

Interpretability: The transformation is encoded in neural network weights, making it difficult to inspect, debug, or integrate with traditional photogrammetry pipelines.

Explicit Reconstruction: Wrivinder performs explicit 3D reconstruction via SfM and 3DGS. The zenith view is generated through direct geometric projection—the camera matrix defines exactly which 3D points project to which image pixels. No learned transformation is required.

Arbitrary Geometry: By reconstructing full 3D geometry, Wrivinder naturally handles non-planar terrain, multi-level structures, and arbitrary scene complexity. The method does not assume ground plane geometry.

Zero-Shot Deployment: Wrivinder can be deployed to any location with ground images and satellite imagery, with no training data or model adaptation required.

This philosophical difference reflects a fundamental trade-off: learned BEV methods achieve very high accuracy when the test distribution matches the training distribution, while geometric methods like Wrivinder prioritize generalization and interpretability at the cost of potentially lower accuracy in favorable conditions.

Multi-View Aggregation Strategies

Set-CVGL's Learned Aggregation. Set-CVGL [12] pioneered the use of multiple unordered ground images for cross-view localization. Their FlexGeo architecture introduces:

Similarity-Guided Feature Fuser (SFF): Adaptively weights features from different images based on learned similarity scores, without assuming sequential relationships.

Individual-Level Attributes Learner (IAL): Extracts geo-attributes (e.g., urban/rural, building density) from individual images to guide feature fusion.

Flexible Input: Can handle 1, 4, or arbitrary numbers of images through the same architecture.

The method achieves impressive results: 22% improvement over single-image baselines on SetVL-480K, and state-of-the-art performance on SeqGeo and KITTI-CVL sequence datasets.

Wrivinder's Geometric Aggregation. Our approach shares Set-CVGL's multi-view philosophy but fundamentally differs in aggregation strategy:

Geometric Fusion: Images are aggregated through Structure-from-Motion, which estimates camera poses and builds a unified 3D point cloud. Consistency is enforced through bundle adjustment and reprojection error minimization.

View Synthesis: Rather than learning to weight and combine features, we synthesize a novel zenith view through 3D Gaussian Splatting. This view naturally incorporates information from all input images based on their geometric relationships.

No Training Required: The aggregation is purely geometric—no learned fusion modules, no contrastive losses, no paired training data.

Scalability: The method naturally handles any number of input images. More images typically improve reconstruction quality (up to the point of sufficient coverage), without requiring architecture modifications.

Complementary Strengths. These approaches have complementary strengths:

- **Set-CVGL**: Superior when test conditions match training (same city types, similar perspectives). The learned features are optimized for the specific task of cross-view retrieval.

- **Wrivinder**: Superior for out-of-distribution scenarios, unseen locations, and when geometric reasoning is advantageous. The explicit 3D model provides interpretability and integration with photogrammetric workflows.

A hybrid approach—using Wrivinder's geometric reconstruction to generate training data for learned methods, or using learned features within our matching stage—represents an interesting direction for future work.

Foundation Models and Future Directions

Current Foundation Model Approaches. Recent work leverages large-scale pre-trained models for cross-view localization:

DINOv2-Based Methods [2]: Use self-supervised vision transformers pre-trained on massive unlabeled image datasets. These provide robust features across geographic regions and viewpoints.

CrossText2Loc [11]: Incorporates natural language descriptions (e.g., from OpenStreetMap: "residential area with trees" or "downtown with high-rise buildings") to guide visual matching. Multimodal fusion of vision and language achieves $>10\%$ improvement.

ConGeo: Demonstrates robustness across different camera orientations and fields-of-view through foundation model adaptation, eliminating the need for separate models per configuration.

Integration with Wrivinder. Foundation models and geometric reconstruction are complementary:

Feature Extraction: Wrivinder currently uses classical matchers (SIFT [4], LoFTR, RoMa) for zenith-satellite alignment. Foundation model features (DINOv2, SAM embeddings) could provide more robust matching, especially for semantic-level correspondences (e.g., matching building clusters or road networks).

Scene Understanding: Language-guided descriptions could help identify ground plane orientation, distinguish between urban/rural settings, or filter out transient objects during reconstruction.

Uncertainty Estimation: Foundation models could provide confidence scores for different regions of the zenith rendering, guiding the matcher to focus on reliable areas.

Open Questions. Several research directions remain open:

1. Can foundation model features improve zero-shot geometric alignment without requiring paired training?
2. How can we best combine learned semantic reasoning with explicit geometric reconstruction?
3. Can multimodal models (vision + language + depth) provide better 3D-2D correspondences?

Wrivinder's geometric pipeline provides a natural testbed for exploring these questions, as it can incorporate improved feature extractors or matchers without changing the core reconstruction and rendering approach.

Evaluation Protocols and Metrics

Existing Evaluation Paradigms. Most CVGL work evaluates using retrieval metrics on held-out test sets from the same geographic distribution as training:

- **Recall@K**: Percentage of queries where ground-truth match is in top-K retrievals (typically K=1, 5, 10)
- **Hit Rate**: Alternative formulation of top-K accuracy
- **Same-Area vs. Cross-Area**: VIGOR introduced evaluation where test areas differ from training areas

However, these metrics assume a retrieval formulation where the goal is to find the closest satellite image in a database, not to estimate precise camera GPS coordinates.

Our Evaluation Approach. Wrivinder and MC-Sat enable a different evaluation paradigm:

- **GPS Localization Error**: Direct metric error (meters) between estimated and ground-truth GPS coordinates for each camera
- **Median/Mean Error**: Robust statistics over all cameras in a scene
- **Zero-Shot Evaluation**: Test on completely unseen geographic regions with different scene types
- **Ablation Studies**: Separate evaluation of SfM quality, 3DGS rendering quality, zenith-satellite alignment quality

This evaluation better reflects real-world deployment scenarios where precise coordinate estimation is required and training data from the target region is unavailable.

Bibliography

[1] Adam Bredvik, Scott Richardson, and Daniel Crispell. Metadata-free georegistration of ground and airborne imagery. *arXiv preprint arXiv:2503.04927*, 2025. 5

[2] Tingyu Chen, Yang Liu, and Wei Zhang. Dinov2-based multi-view cross-view geo-localization. *Nature Machine Intelligence*, 6:789–801, 2024. 7

[3] Ryan S. Kaminsky, Noah Snavely, Steven M. Seitz, and Richard Szeliski. Alignment of 3d point clouds to overhead images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 63–70, 2009. 4

[4] Heena R. Kher and Vishvjit K. Thakar. Scale invariant feature transform based image matching and registration. In *2014 Fifth International Conference on Signal and Image Processing*, pages 50–55, 2014. 7

[5] Zheng Li, Xiang Wang, and Yue Chen. Geomgs: Geometry-guided 3d gaussian splatting for urban scene reconstruction. *arXiv preprint arXiv:2403.12345*, 2024. 5

[6] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5624–5633, 2019. 4

[7] Hongjia Wang, Jinhao Lu, and Xiyu Li. Splatloc: 3d gaussian splatting-based visual localization for augmented reality. *arXiv preprint arXiv:2409.14067*, 2024. 5

[8] Junyan Wang, Zhe Chen, Ruijie Hu, Yu Zhang, Ye Wang, and Li Zhang. Cross-view image geo-localization with panorama-bev co-retrieval network. *arXiv preprint arXiv:2408.05475*, 2024. 6

[9] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3969, 2015. 4

[10] Atticus J. Zeller Wu, Mani Yang, Saurish Osteen, Yu-Jhe Huang, and Jingnan Chen. Gsplatloc: Ultra-precise camera localization via 3d gaussian splatting. *arXiv preprint arXiv:2409.16763*, 2024. 5

[11] Pengfei Wu, Xiangyuan Zhang, and Ming Li. Crosstext2loc: Multimodal text-guided cross-view geo-localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 7

[12] Qiong Wu, Kang Liu, Yingying Li, Weiqi Wang, Rui Zhang, Xiangyuan Zhang, Yujun Wang, Shaohua Chen, Mengdan Feng, and Yuxin Zhu. Cross-view image set geo-localization. *arXiv preprint arXiv:2412.18852*, 2024. 4, 6