X-GeoP2P: Adapting Cross-View Retrieval and Georeferenced Reconstruction for the 2026 WRIVA Challenge

Wyatt Petula
Pennsylvania State University
wpp5085@psu.edu

Soumik Ghosh
University of Central Florida
Soumik.Ghosh@ucf.edu

Qingyang Liu
University of Central Florida
qingyang.liu2@ucf.edu

David Shatwell
University of Central Florida
david.shatwell@ucf.edu

Mubarak Shah
University of Central Florida
shah@crcv.ucf.edu

Abstract

The goal of cross-view geolocalization is to estimate the geographic location of a ground-level image by matching it to aerial or satellite imagery. This task is difficult because ground and overhead views differ in perspective, scale, and appearance. The 2026 WRIVA Cross-View Geo-Localization Challenge centers on off-road ground-image sequences that must be localized within large satellite images [14]. Its data differ from the road-centered video used to train many existing cross-view geolocalization systems.

We present X-GeoP2P, a coarse-to-fine localization pipeline that combines adapted versions of GAReT and X-VGGT. This report focuses mainly on my contribution to the pipeline: the GAReT adaptation, denoted as GAReT-ω, and its data preprocessing and postprocessing. I describe the baseline models, their adaptations, the combined pipeline, and their performance on the challenge.

1. Introduction

Determining where an image was captured is a fundamental problem in computer vision with applications in autonomous navigation, robotics, mapping, disaster response, geospatial intelligence, and GPS-denied environments. Visual place recognition matches a query against a geotagged reference captured from a similar viewpoint. In some situations, such reference imagery may be unavailable, geographically incomplete, or outdated. Cross-view geolocalization addresses this limitation by matching a ground-view query to georeferenced overhead imagery, which is available over a substantially larger fraction of the Earth's surface.

Cross-view geolocalization is inherently difficult. A ground image primarily captures facades, vegetation, elevation, perspective-view distortion, and occluded objects and structures. Conversely, a satellite image captures rooftops, road networks, and parcel geometry in an orthographic setting. Even when a query and reference image correspond to the same location, determining the relationship between their views can be challenging due to differences in orientation, scale, field of view, illumination, season, weather, and acquisition year. Furthermore, dynamic objects and occlusions can remove or alter visual cues, while patterns like city grids can produce hard negatives that are visually plausible but geographically incorrect.

This project's contributions are as follows:

1. We benchmark a baseline implementation of GAReT on the challenge to evaluate its performance under a large domain shift.

2. We introduce GAReT-ω and X-VGGT-ω, adaptations of GAReT and X-VGGT for the challenge.

3. We propose X-GeoP2P, a coarse-to-fine pipeline that uses GAReT-ω to extend X-VGGT-ω to kilometer-scale satellite imagery.

2. Related Work

Early methods of cross-view geolocalization relied on handcrafted descriptors or geometric transformations designed to reduce the discrepancy between ground and overhead views (i.e., polar coordinates, panoramas). More recent methods use deep neural networks, metric-learning objectives, attention, and transformer architectures to learn cross-view representations [2, 3]. These methods can be organized along two orthogonal axes: whether location is represented via discretization or a continuous pose, and whether cross-view alignment is learned primarily in 2D or 3D.

In the first category, retrieval-based methods embed data into a shared feature space and return the center coordinate of a reference image or patch. GAReT extends this discretized formulation to ordered videos by combining sequence-level feature aggregation with autoregressive frame-to-patch selection [2]. Recent retrieval systems have also expanded beyond conventional ground-to-satellite image pairs. CrossText2Loc retrieves satellite references from natural-language scene descriptions [5], while MM-Geo composes UAV imagery with complementary depth, point-cloud, or textual information before retrieval against a satellite reference map [6].

Fine-grained 2D methods estimate a continuous location, and often a heading, within a local aerial reference. FG^2 uses fine-grained feature matching for this setting [4]. Towards Unconstrained Cross-View Pose Estimation uses a learned pose-querying procedure while relaxing common assumptions about field of view, camera roll and pitch, horizon alignment, and image projection [7]. GeoFlow performs real-time 2-DoF localization by iteratively refining a population of candidate location hypotheses [8], whereas RHO predicts 3-DoF camera pose from panoramic imagery using OpenStreetMap as the overhead reference rather than satellite imagery [9]. These methods provide finer predictions than gallery retrieval.

A third line of work incorporates learned or explicit 3D geometry. X-VGGT jointly predicts multi-view camera poses, dense 3D point maps, and a scene-level transformation into a georeferenced satellite frame [12], building on VGGT [11]. Similarly, Cross3R reconstructs cross-view geometry and camera poses from combinations of satellite, UAV, and ground imagery [13]. Geo^2 uses priors from geometric foundation models to align ground and aerial features in a shared 3D-aware latent space [10]. These approaches move beyond patch-based prediction by grounding 3D scene reconstructions and camera poses continuously within the reference map.

This grouping shows a practical trade-off. Retrieval methods can search large aerial databases, but their outputs are coarse. Continuous 2D and 3D methods are more precise but usually operate on smaller regions. X-GeoP2P combines these approaches: GAReT-ω first reduces the large challenge references to a localized crop, which X-VGGT-ω uses for continuous per-frame pose predictions.

3. Problem Formulation and Data

3.1. Evaluation Metric

Let the WRIVA Challenge contain sites $S = \{s_i\}_{i=1}^{N}$. Site s_i contains ground-view frames $F_i = \{f_{ij}\}_{j=1}^{M_i}$ and one or more georeferenced satellite references $R_i = \{r_{ik}\}_{k=1}^{O_i}$. Each frame has an unknown camera coordinate $\mathbf{y}_{ij} = (\mathrm{lat}_{ij}, \mathrm{lon}_{ij})$, and the goal is to predict $\hat{\mathbf{y}}_{ij}$. Submission scores are calculated as the average of all horizontal prediction distance errors in meters over all sites:

$$E_{\mathrm{mean}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{M_i}\sum_{j=1}^{M_i} d_{\mathrm{geo}}(\hat{\mathbf{y}}_{ij}, \mathbf{y}_{ij})\right) \quad (1)$$

3.2. WRIVA Challenge Dataset

There is no trivial function to map the challenge dataset into the format GAReT expects, hence the need for GAReT-ω implementations.

The raw challenge dataset contains 149 training sites, three validation sites, and 12 test sites. The number of reference images varies substantially by site. For the training set, the minimum is one, the median is seven, the 75th percentile is 24, the 90th percentile is 58, and the maximum is 114. The validation sites contain between four and 26 references.

Not all sites or references were suitable for fine-tuning. We excluded sites with incorrect satellite coverage, missing data, white-border corruption, or heavy cloud cover. We also excluded sequences with frames outside the selected reference. The remaining 106 training sites and all three validation sites were used for evaluation.

4. Method

4.1. Pipeline Overview

X-GeoP2P consists of four stages, depicted in Figure 1:

1. Convert WRIVA Challenge ground sequences, Maxar references, and metadata into a GAReT-compatible representation.
2. Use GAReT-ω to retrieve an aerial patch for each frame.
3. Aggregate the retrieved patches into a centroid crop for bursts of fixed-length sequences.
4. Run X-VGGT on the ground images and the burst crops to predict continuous camera locations.

4.2. Baseline GAReT

Table 1 shows the different data formats and capture conditions between GAMa and the WRIVA Challenge. To reconcile these differences, we built a baseline WRIVA implementation that keeps GAReT's main inference stages with minimal changes [1, 2].

4.2.1 Ground-View Sequence Preparation

WRIVA Challenge ground images are grouped into fixed-length sequences. The baseline uses eight frames per sequence to match GAReT. Each frame is center-cropped and resized to the landscape resolution expected by the ground encoder.

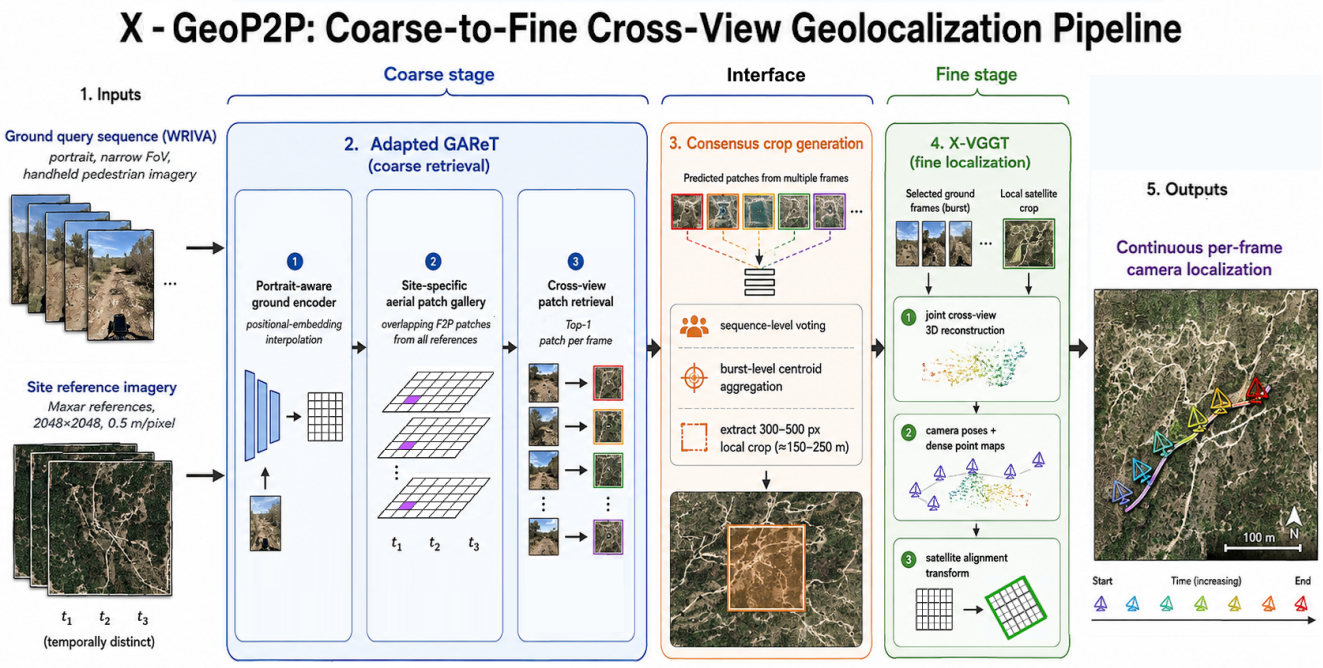

Figure 1. Proposed coarse-to-fine X-GeoP2P pipeline. WRIVA ground-view frames and Maxar reference imagery are processed by GAReT-ω for frame-to-patch retrieval. Sequence-level voting and burst-level centroid aggregation produce a localized satellite crop, which is passed to X-VGGT for continuous per-frame localization.

4.2.2 Aerial Gallery Construction

Every Maxar reference image is treated as an independent candidate source. Each reference is divided into 680 $\times$ 680 px chips, covering 340 m, and resized with bicubic interpolation to 1792 $\times$ 1792 px. This keeps the WRIVA chip footprint geometrically simple while roughly matching GAReT's expected spatial resolution.

All chips from all references for an individual site are placed in a site-specific S2I gallery. Cross-site candidates are not included because the challenge provides site identity.

4.2.3 Sequence-to-Image Retrieval

For each eight-frame query sequence, the ground encoder and GeoAdapter produce a pooled video embedding. The aerial encoder and GeoAdapter produce an embedding for each large aerial chip. Nearest-neighbor retrieval returns the top-10 S2I candidates. The baseline forwards only the top-ranked large chip to the next stage to limit the size of the F2P gallery.

4.2.4 Frame-to-Patch Retrieval

The selected large chip is divided into non-overlapping 256 $\times$ 256 px F2P patches. GeoAdapter is removed, and the image encoders independently embed every query frame and aerial patch. For each frame, the ten nearest F2P candidates are retained.

4.2.5 TransRetriever

For each sequence, TransRetriever receives a similarity matrix and latitude-longitude arrays converted into UTM for the top-k F2P candidates. TransRetriever then autoregressively selects one candidate per frame. The center coordinates of the selected patches become the submitted frame predictions.

4.3. GAReT-ω

GAReT-ω changes the baseline in five main ways: portrait-preserving query encoding, data filtering, multi-positive fine-tuning, deterministic central S2I cropping, and overlapping F2P patches. For standalone GAReT-ω submissions, the top-ranked F2P patch center is used directly. In the combined pipeline, the patch predictions are post-processed and supplied to X-VGGT-ω rather than TransRetriever.

4.3.1 Portrait Positional-Embedding Interpolation

GAReT's ground transformer was pretrained on landscape images, while WRIVA frames are portrait-oriented. Center-cropping removes part of the frame, and direct resizing

Table 1. Relevant differences between the GAReT training domain and the WRIVA Challenge domain.

Property	GAMa / GAReT	WRIVA Challenge
Ground imagery	Vehicle-mounted video	Handheld pedestrian video
Ground orientation	Landscape, approximately horizon-fixed	Portrait, variable camera angle
Prepared model input (W×H)	384 × 216 px	3024 × 4032 px
Reference imagery	1792 × 1792 px, approximately 0.238 m/px	2048 × 2048 px, 0.5 m/px
Scene content	Road-centered driving routes	Mostly off-road locations
Positive mapping	One positive aerial patch per query	Several temporal references may provide valid positives
Supervision	Matching aerial patch or GPS	Per-frame camera GPS
Evaluation metric	Retrieval and distance-threshold recall	Mean camera-position error

changes the scene geometry.

The adapted encoder keeps the portrait layout. Because the pretrained positional embeddings were learned for a different spatial arrangement, their spatial component is interpolated to match the portrait token layout. The special-token embeddings are left unchanged.

This keeps the full field of view and reuses the pretrained transformer weights. It also changes the positional-embedding distribution, and the effect of this change was not measured separately in the final submissions.

4.3.2 Data Filtering and Reference Selection

For multi-positive reference experiments, all filtered Maxar references are used per site. For single-reference experiments, one Maxar reference is retained per site. Candidate selection criteria include the lowest off-nadir angle and the acquisition month closest to the ground-image capture month. The best recorded GAReT-ω configuration used the nearest-month criterion. Selecting one reference reduces temporal ambiguity, whereas selecting the lowest nadir angle (angular displacement from BEV) reduces occlusion caused by tall buildings.

4.3.3 Deterministic S2I Region

In the WRIVA training and validation data, ground trajectories are concentrated near the center of sites. Learned S2I retrieval is therefore not necessary than in GAMa, where the gallery contains many spatially distinct large aerial regions. An imperfectly adapted GeoAdapter can introduce a large coarse error before F2P retrieval.

GAReT-ω replaces multiple learned S2I candidates with one 768 × 768 center crop from the selected reference image. This crop acts as a deterministic S2I region. The S2I search and GeoAdapter ranking are bypassed, while the image encoders are retained for F2P retrieval inside the central region. This design intentionally uses the challenge's coarse location prior and focuses model capacity on the frame-level cross-view match.

4.3.4 Overlapping F2P Patches

GAReT-ω uses the same patch size as the baseline but with 50% stride, as overlap reduces the worst-case distance from a ground-truth position to the nearest patch center. It also creates a denser local neighborhood for X-VGGT crop preparation. This is important because adjacent WRIVA frames may be separated by only a few meters, whereas non-overlapping patch centers can be much farther apart.

4.3.5 Multi-Positive Bidirectional Fine-Tuning

For multi-positive experiments, GAReT-ω pairs each ground image with multiple F2P patches across references, as spatially-identical patches from different dates are all valid positives for the same query frame.

The adapted training data are constructed at the frame-patch-pair level. For each eligible frame and sampled reference image, the frame is paired with the F2P patch containing its ground-truth location. If multiple references are used, the same frame can appear in several examples, each paired with the spatially corresponding patch from a different acquisition. These pairs are marked as equivalent. When calculating the loss, all patches in the same positive group are treated as positives and are excluded from the negative set. This avoids penalizing time-invariant matching.

Fine-tuning starts from the released GAReT image-model checkpoint and runs for ten epochs with a batch size of 64 and a cosine learning-rate schedule.

4.3.6 Top-1 F2P Prediction

X-VGGT-ω replaces TransRetriever as the final localizer. We retain the most similar F2P patch per frame for crop aggregation instead of passing a larger candidate set downstream.

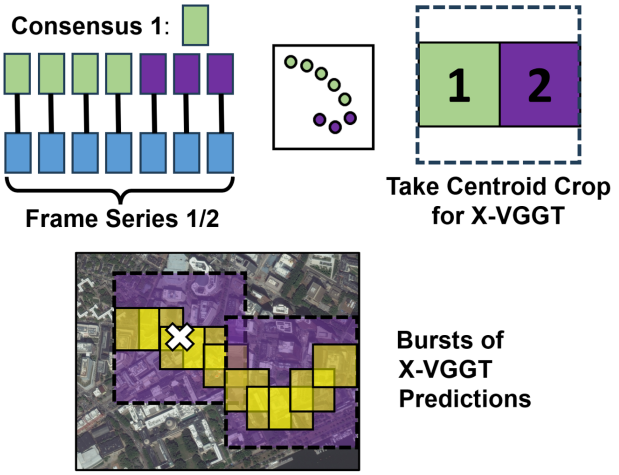

Figure 2. Consensus aggregation from frame-level retrievals to an X-VGGT crop. Coordinates remain in the selected reference image until the final geospatial conversion.

4.4. Consensus Crop Preparation for X-VGGT

F2P predictions can vary between nearby or visually similar patches. To create a stable input for X-VGGT, we aggregate them at the sequence and burst levels. A visualization of this process is depicted in Figure 2.

4.4.1 Sequence-Level Voting

For each sequence, every frame votes for its top-ranked F2P patch. The patch with the most votes becomes the sequence prediction. Ties are broken by mean retrieval similarity and then by patch identifier.

4.4.2 Burst-Level Centroid

Consecutive sequences are grouped into bursts. If the center of the consensus patch for sequence i is (x_i, y_i), the burst center is

$$(\bar{x}, \bar{y}) = \left(\frac{1}{B} \sum_{i=1}^{B} x_i, \frac{1}{B} \sum_{i=1}^{B} y_i \right), \tag{2}$$

where B is the number of sequences in the burst. A square crop is extracted from the selected reference image centered at $(\bar{x}, \bar{y})$. Crops extending outside the reference boundaries are padded with white.

4.4.3 Crop Scales

We perform experiments using 300-pixel and 500-pixel crops. At 0.5 m per pixel, these correspond to 150 m and 250 m.

4.5. X-VGGT-ω and Inference

X-VGGT-ω was fine-tuned using 600-pixel center crops, corresponding to 300 m. It also uses trajectory outlier detection to correct noisy predictions.

5. Experimental Setup

5.1. Evaluated Configurations

- **Baseline GAReT, every frame:** baseline WRIVA reformatting, released GAReT checkpoints, and a prediction for each frame. No fine-tuning.
- **Baseline GAReT, every fifth frame:** the same model, but predicting every fifth frame and interpolating the intermediate positions, following GAReT's temporal sampling more closely. No fine-tuning.
- **GAReT-ω, every frame:** portrait positional interpolation, one nearest-month reference per site, a 768-pixel central S2I region, 256-pixel F2P patches with 50% stride, multi-positive fine-tuning, and eight-frame sequences. Multi-positive fine-tuning on same crop and patch setup.
- **GAReT-ω, every fifth frame:** the adapted model with position interpolation between every-fifth-frame predictions. Multi-positive fine-tuning on same crop and patch setup.
- **Baseline X-VGGT:** the original model applied to the WRIVA Challenge setting without the final adaptation. No fine-tuning.
- **X-VGGT-ω:** fine-tuning on 600-pixel center crops, 600-pixel inference crops, and trajectory outlier correction. Fine-tuning on 300 m center crops.
- **X-GeoP2P:** five experiments involving five-and-two-sequence bursts at 150 m and 250 m, plus one burst containing all sequences.

6. Results

6.1. Challenge Scores

Table 2 reports the recorded challenge scores. Lower mean horizontal error is better.

The every-frame GAReT-ω experiment reduced the error from 518.57 m to 198.93 m, a 61.6% reduction. In the every-fifth-frame setting, the error decreased from 481.38 m to 195.30 m, a 59.4% reduction. This shows that the baseline did not transfer well to WRIVA and that the full set of adaptations helped.

X-VGGT-ω reduced the error from 135.49 m to 65.92 m. This component result used a center crop, so it is not directly comparable to the combined pipeline, which uses a GAReT-predicted crop.

Table 2. Recorded 2026 WRIVA Challenge mean horizontal error.

Configuration	Mean Error (m)↓
Individual Model Experiments	
Baseline GAReT (every frame)	518.57
Baseline GAReT (every 5th frame + interpolation)	481.38
GAReT-ω (every frame)	198.93
GAReT-ω (every 5th frame + interpolation)	195.30
Baseline X-VGGT	135.49
X-VGGT-ω (so far)	**65.92**
X-GeoP2P Experiments	
2 sequences, 500-pixel crop	156.97
5 sequences, 500-pixel crop	153.76
All sequences, 500-pixel crop	128.91
2 sequences, 300-pixel crop	128.88
5 sequences, 300-pixel crop	**125.09**

6.2. X-GeoP2P Challenge Results

The five combined configurations achieved errors between 125.09 m and 156.97 m. The best result was 125.09 m using five-sequence bursts and 300-pixel crops. This improved on the every-frame GAReT-ω result by 73.84 m, or 37.1%, showing that X-VGGT-ω added useful local refinement when the retrieved crop contained enough relevant context.

Crop size had a larger effect than burst length in these experiments. Changing from a 500-pixel crop to a 300-pixel crop reduced error by 28.09 m for two-sequence bursts and by 28.67 m for five-sequence bursts. Increasing the burst length from two to five sequences produced smaller gains: 3.21 m with 500-pixel crops and 3.79 m with 300-pixel crops. The all-sequence, 500-pixel configuration reached 128.91 m, which was better than the shorter 500-pixel settings but slightly worse than either 300-pixel setting.

The best combined result remained 59.17 m worse than the 65.92 m center-crop X-VGGT-ω result. Since the center-crop experiment avoids GAReT crop error, this gap suggests that coarse crop placement and aggregation were the main limits of the combined pipeline.

6.3. Effect of Temporal Subsampling and Interpolation

For baseline GAReT, predicting every fifth frame and interpolating the intermediate positions improved the score by 37.19 m, or 7.2%. The every-frame baseline was likely noisy, and interpolation smoothed part of that noise.

For GAReT-ω, the same procedure improved the score by only 3.63 m, or 1.8%. This smaller change suggests that the adapted per-frame predictions were already more stable.

Table 3. Minimum square crop size required to contain ground-truth trajectories in the training set. Values are given in pixels of the original reference image.

Grouping	P50	P75	P90	P95	Max
All sequences	381.0	543.0	712.0	802.4	1001
2-sequence burst	21.0	37.0	85.0	131.0	888
5-sequence burst	54.0	103.0	175.7	238.6	888

6.4. Ground-Trajectory Crop Geometry

Before selecting burst crop sizes, the project measured the minimum square crop required to contain the ground-truth frame positions for different temporal groupings. Table 3 reports the training-set distribution. Pixel values can be converted to meters by multiplying by 0.5.

Most local bursts are compact. A 300-pixel crop is larger than the 95th-percentile ground-truth extent for both two- and five-sequence bursts. A 500-pixel crop gives more room for GAReT error. The large maximum values show that a small number of bursts are much wider than the rest.

For all sequences in a site, a 500-pixel crop is smaller than the 75th-percentile extent of 543 pixels. It cannot contain every frame for many sites, even with a correct center. The all-sequence experiment tests whether the extra views still help when some frames fall outside the crop.

6.5. Interpretation of GAReT-ω Results

The best GAReT-ω configuration changes several parts of the system at once, including preprocessing, filtering, reference selection, training, and patch generation. The final score shows that the full configuration helped, but it does not show which change mattered most. Controlled ablations are needed.

GAReT predicts a patch center, so its output is still coarse. X-VGGT predicts continuous camera locations from multiple images. The combined results show that X-VGGT can improve on GAReT, but the quality of the predicted crop currently hinders the final accuracy.

7. Discussion

7.1. Why the Baseline Transferred Poorly

Baseline GAReT differs from WRIVA in several ways. Its center crop removes part of the portrait image, its encoders were trained mainly on road scenes, and its non-overlapping F2P grid gives coarse patch centers. The S2I stage can also choose the wrong region. If that happens, later stages cannot recover the correct patch.

7.2. Benefits and Risks of a Single Reference

Selecting one reference per site simplifies the pipeline and keeps every patch center in the same pixel coordinate system. Nearest-month selection may also reduce seasonal differences.

However, the nearest month is not always the best-looking reference. Clouds, shadows, viewing angle, vegetation, and image quality can matter more. A future version should test multiple references during inference.

7.3. Consensus Aggregation Trade-Offs

Sequence voting removes some isolated frame errors, but it can also repeat a shared mistake. It also treats nearby patches as different votes even when they describe almost the same area.

A better method could cluster patch centers or weight them by retrieval similarity. We used majority voting and a centroid because they are simple and reproducible.

7.4. Crop-Size Trade-Off

A small crop keeps X-VGGT close to its training distribution, but it may miss the true location when GAReT is inaccurate. A larger crop is more likely to contain the trajectory but includes more distractors. In both matched burst-size comparisons, the 300-pixel crop performed better than the 500-pixel crop.

A future system could choose the crop size from the spread or confidence of the GAReT predictions instead of using one fixed size.

7.5. Attribution and Reproducibility

GAReT-ω and X-VGGT-ω were developed by different contributors: me and my mentors, respectively. The code for X-GeoP2P will be available online later this summer.

8. Limitations

The project has several limitations.

First, the GAReT-ω result combines several changes, so the effect of each change is unknown.

Second, mean horizontal error does not show the median, tail errors, site-level performance, or how often the true trajectory falls inside the X-VGGT-ω crop.

Third, the combined pipeline did not preserve the center-crop accuracy of X-VGGT-ωThe current experiments do not separate crop-placement error from X-VGGT-ω localization error.

Fourth, the system assumes that the site is known and that a usable georeferenced satellite image is available. It may fail when the selected reference is outdated, cloudy, misregistered, or does not cover the query.

Fifth, F2P predictions are patch-center approximations. Overlap reduces this error but increases the gallery size. Top-1 selection also discards uncertainty.

Finally, some details of X-VGGT-ω were not available to me at the time of writing this report.

9. Future Work

I learned a lot from developing X-GeoP2P this summer; however, if I want my contributions to be meaningful, I feel strongly that I will need to start over. If I could time travel back to the start of the program, my first step would be to perform a more comprehensive literature review of cross-view geolocalization. At the beginning of the program, my cohort was encouraged not to read the literature because it would dilute our novelty and hinder our ability to quickly produce results. Although this approach has merits, it is challenging to do good research when you do not know the state of the field and what resources are available to build on. When possible, I think research should be performed using a top-down learning approach, not bottom-up.

I would also step back from the WRIVA Challenge specifics and instead investigate its core problem, at least for the first week or two. My understanding of computer vision research is that it should be thorough and principled. This typically involves benchmarking many existing methods, contributing a dataset or methodology that addresses a significant gap better than those methods, and empirically persuading other researchers to trust and care about your findings. Success in these types of challenges is less demanding, as an ad-hoc method that depends on heuristics is no less valuable than a novel contribution. In the end, all that matters is which submission score is lower. The challenge was a good starting point for my introduction to modern computer vision research, but I still have more to learn on my path to becoming a professional researcher.

10. Conclusion

This project tested cross-view geolocalization under a shift from vehicle road imagery to handheld off-road imagery. Baseline GAReT achieved 518.57 m mean horizontal error when predicting every frame. GAReT-ω reduced this to 198.93 m, or 195.30 m when predicting every fifth frame and interpolating.

X-VGGT-ω achieved 65.92 m when given a center crop. The combined X-GeoP2P pipeline used GAReT-ω predictions to generate X-VGGT crops and achieved a best result of 125.09 m with five-sequence bursts and 300-pixel crops. This improved on GAReT-ω but the gap to center-crop X-VGGT-ω shows that coarse crop retrieval needs refinement.

Acknowledgments

This work was completed through the Research Experiences for Undergraduates (REU) program at the University of Central Florida Center for Research in Computer Vision. I thank my mentors for developing and adapting X-VGGT, advising the experimental design, and supporting integration with the GAReT retrieval pipeline. I also thank Dr. Shah for his insight into top research, taking walks, and higher education.

References

[1] S. Vyas, C. Chen, and M. Shah, "GAMa: Cross-view Video Geo-localization," in *European Conference on Computer Vision (ECCV)*, pp. 440–456, 2022. 2

[2] M. S. Pillai, M. N. Rizve, and M. Shah, "GAReT: Cross-view Video Geolocalization with Adapters and Auto-Regressive Transformers," in *European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[3] S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer Is All You Need for Cross-View Image Geo-Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1162–1171, 2022. 1

[4] Z. Xia and A. Alahi, "FG^2: Fine-Grained Cross-View Localization by Fine-Grained Feature Matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6362–6372, 2025. 2

[5] J. Ye, H. Lin, L. Ou, D. Chen, Z. Wang, Q. Zhu, C. He, and W. Li, "Where am I? Cross-View Geo-localization with Natural Language Descriptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5890–5900, 2025. 2

[6] Y. Ji, B. He, Z. Tan, and L. Wu, "MMGeo: Multimodal Compositional Geo-Localization for UAVs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25165–25175, 2025. 2

[7] A. Wollam, K. Ashley, M. Shugaev, O. Arend, I. Semenov, H. Dashtestani, S. Ravi, and N. Jacobs, "Towards Unconstrained Cross-View Pose Estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8448–8457, 2026. 2

[8] A. Abu Lehyeh, X. Zhang, A. Arrabi, W. Sultani, C. Chen, and S. Wshah, "GeoFlow: Real-Time Fine-Grained Cross-View Geolocalization via Iterative Flow Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2

[9] J. Zheng, R. Dai, R. Liu, Z. Zeng, Y. Chen, F. Wang, K. Peng, K. Yang, J. Zhang, and R. Stiefelhagen, "RHO: Robust Holistic OSM-Based Metric Cross-View Geo-Localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33727–33737, 2026. 2

[10] Y. Zhang, X. Zhang, G. Sun, Z. Lyu, S. Wshah, and C. Chen, "Geo^2: Geometry-Guided Cross-View Geo-Localization and Image Synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19432–19442, 2026. 2

[11] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual Geometry Grounded Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025. 2

[12] D. G. Shatwell and M. Shah, "Georeferenced Cross-View 3D Reconstruction from Ground and Satellite Images," manuscript under review, 2026. *Bibliographic year and venue to be updated upon submission or acceptance.* 2

[13] Q. Wang, Z. Tuo, X. Ze, and Y. Shi, "Seeing Across Skies and Streets: Feedforward 3D Reconstruction from Satellite, Drone, and Ground Images," *arXiv preprint arXiv:2605.07978*, 2026. 2

[14] Intelligence Advanced Research Projects Activity, "IARPA Announces the 2026 WRIVA Cross-View Geo-Localization (CVGL) Challenge," Apr. 17, 2026. Available through the IARPA newsroom and IEEE DataPort challenge page. 1