From Patches to Points: Adapting GAReT for Temporal Geolocalization and Retrieval-Guided X-VGGT

Wyatt Petula[1], Soumik Ghosh[2], Qingyang Liu[2], David Shatwell[2], Dr. Mubarak Shah[2]

[1]*The Pennsylvania State University,* [2]*University of Central Florida*

Introduction

The 2026 WRIVA Challenge:

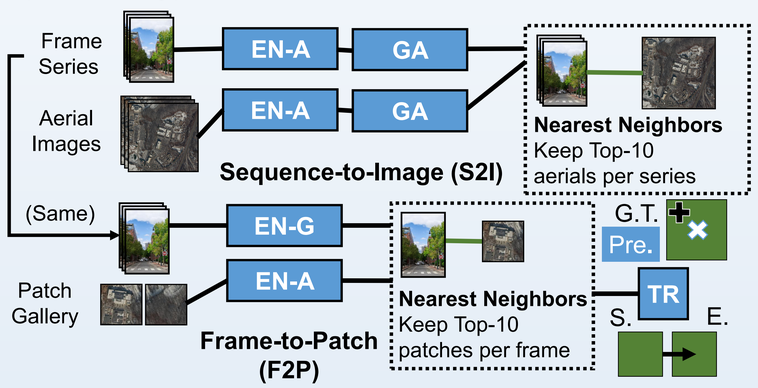
Given…

- A sequence of portrait-view, off-road, ground frames (1)
- A gallery of spatially-identical, temporally-diverse aerial tiles (2)

…Predict the location of each frame (3).

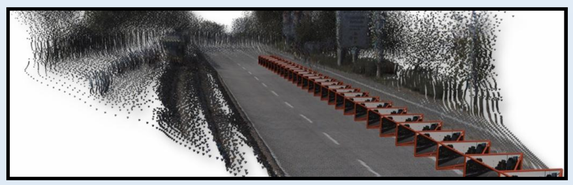


GAReT: Takes street videos and spatially-diverse aerials. Predicts frame positions using patch center coordinates. Trained on GAMa.



EN-A = Aerial Patch Encoder EN-G = Frame Encoder
GA = GeoAdapter TA = TransRetriever

X-VGGT: Takes a 150m–250m satellite image crop, performing best on small crops. Predicts frame positions directly.



X-VGGT predictions for a sequence of frames.

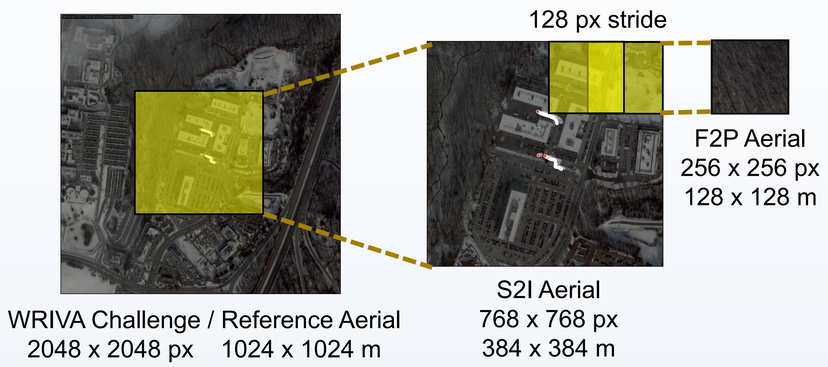
Initial Motivation:

- GAReT leverages sequential relationships that are conducive for the challenge
- GAReT's aerial patches are small and usable for X-VGGT

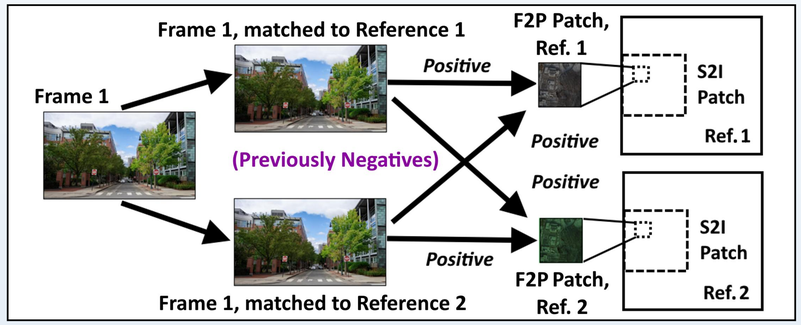
Approach

Current Challenge Adaptation of GAReT:

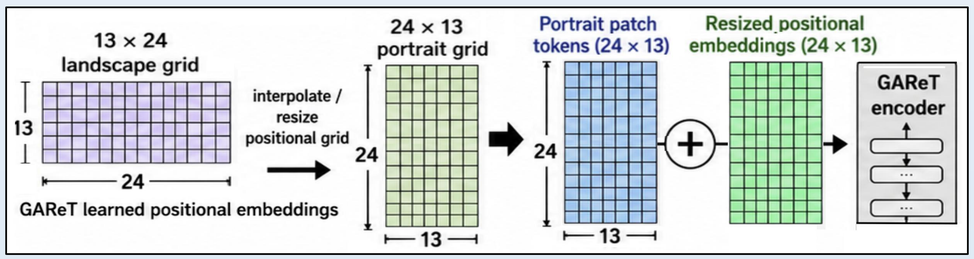
- GeoAdapter is redundant because frames are always near the center of aerial references
- X-VGGT predicts trajectories better than TransRetriever
- **The image encoders are all we need**



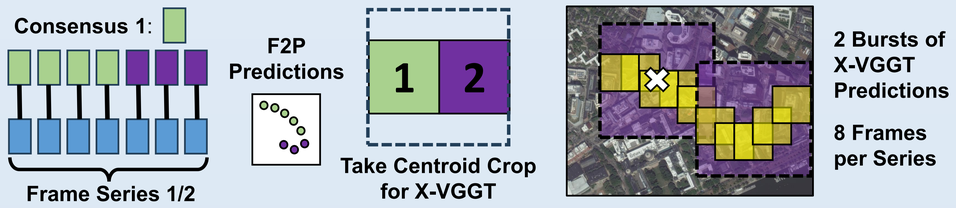
The GAReT-only challenge adaptation. White dots are ground-truth frame positions.



GAReT uses single-positive triplet loss, which is counterproductive for this task. The adaptation uses multi-positive loss.



GAReT takes landscape-view frames. The challenge dataset uses portrait-view. Positional embedding interpolation maximizes information preserved versus center cropping.



Using image encoder predictions to fetch optimal crops for X-VGGT.

Results

Challenge Submission Results:

- Scores are the horizontal distance error of predictions in meters

Approach	Score (m)
GAReT – Predict every frame	518.57
GAReT – Predict every 5th frame, interpolate predictions in-between	481.38
Adapted GAReT	Pending
Adapted GAReT + X-VGGT	Pending

General Observations:

- GAReT specializes in street trajectories and depends on assumptions not supported by off-road trajectories
- Challenge dataset is not orthorectified, offering challenges for approaches like GAReT and new opportunity for approaches like X-VGGT

Future Work

- Seek models like GAReT that specialize in off-road video geolocalization
- Generalize to problem-relevant datasets beyond the challenge
- Investigate novel 3D approaches to the problem

References

[1] M. S. Pillai, M. N. Rizve, and M. Shah, "GAReT: Cross-view video geolocalization with adapters and auto-regressive transformers," in *Computer Vision—ECCV 2024*, Milan, Italy, 2024, pp. 466–483, doi: 10.1007/978-3-031-73030-6_26.
[2] M. S. Pillai, "GAReT: Cross-view video geolocalization with adapters and auto-regressive transformers," *GitHub repository*, 2024. [Online]. Available: https://github.com/manupillai308/GAReT.
[3] S. Vyas, C. Chen, and M. Shah, "GAMa: Cross-view video geo-localization," in *Computer Vision—ECCV 2022*, Tel Aviv, Israel, 2022, pp. 440–456, doi: 10.1007/978-3-031-19836-6_25.
[4] D. G. Shatwell and M. Shah, "Georeferenced cross-view 3D reconstruction from ground and satellite images," unpublished manuscript, Institute of Artificial Intelligence, University of Central Florida, Orlando, FL, USA, 2026.
[5] IEEE DataPort, "2026 WRIVA Cross-View Geo-Localization (CVGL) Challenge," 2026. [Online]. Available: https://ieee-dataport.org/competitions/wriva-cvgl-challenge-2026.

Acknowledgements

I would like to thank the following individuals for their advisement and support this summer:
Dr. Mubarak Shah, **Dr. Niels da Vitoria Lobo**, and **Melody Halbert**.

I would also like to thank the following individuals for their mentorship:
Soumik Ghosh, Qingyang Liu, David Shatwell, and **Zachary Roch.**