From Patches to Points: Adapting GAReT for the 2026 WRIVA Challenge and Retrieval-Guided X-VGGT

Student: Wyatt Petula

Mentors: Soumik Ghosh, Qingyang Liu, Dr. Mubarak Shah

The WRIVA Challenge is a location guessing game



(1)



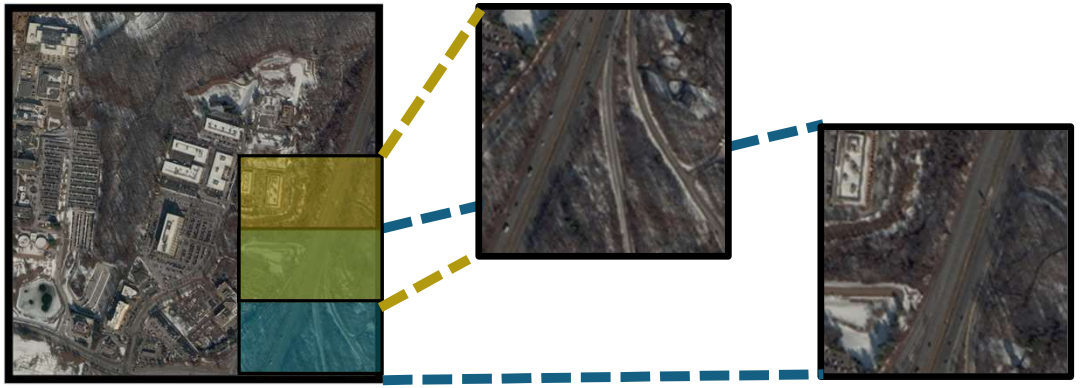
(2)



(3)

GAReT narrows the search space; X-VGGT does fine localization

GAReT:



X-VGGT:



Adapted from Georeferenced Cross-View 3D Reconstruction from Ground and Satellite Images (preprint)

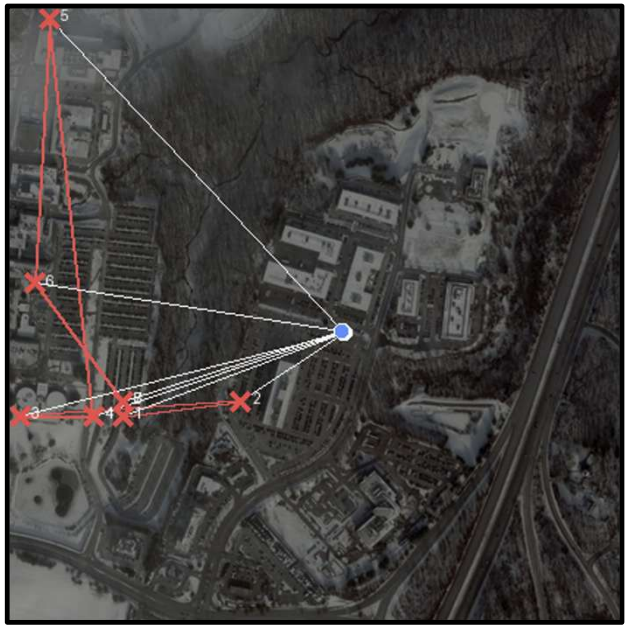
I adapted GAReT to the challenge's dataset and geolocalization subproblem



GAReT Predictions (Street)

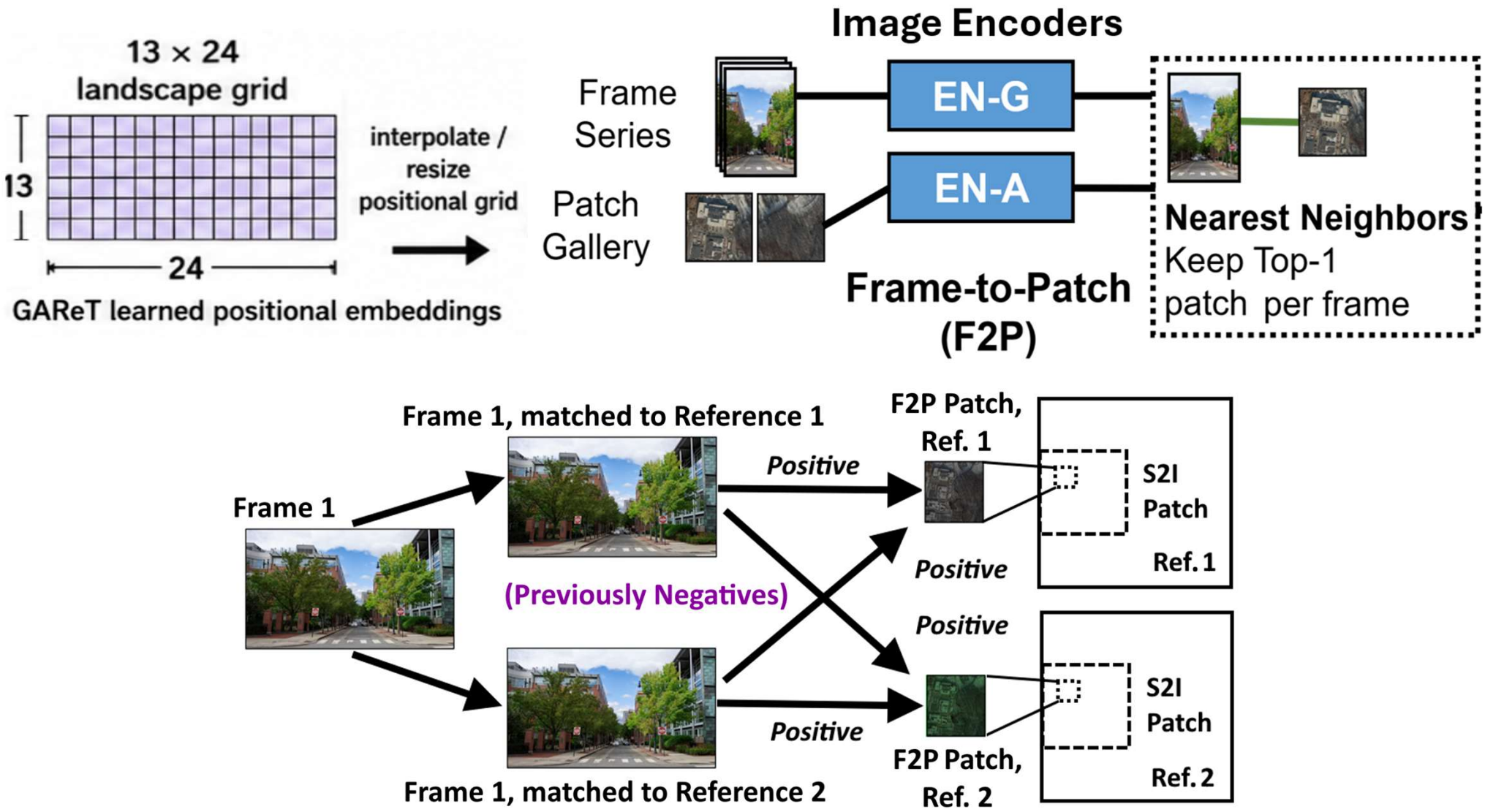


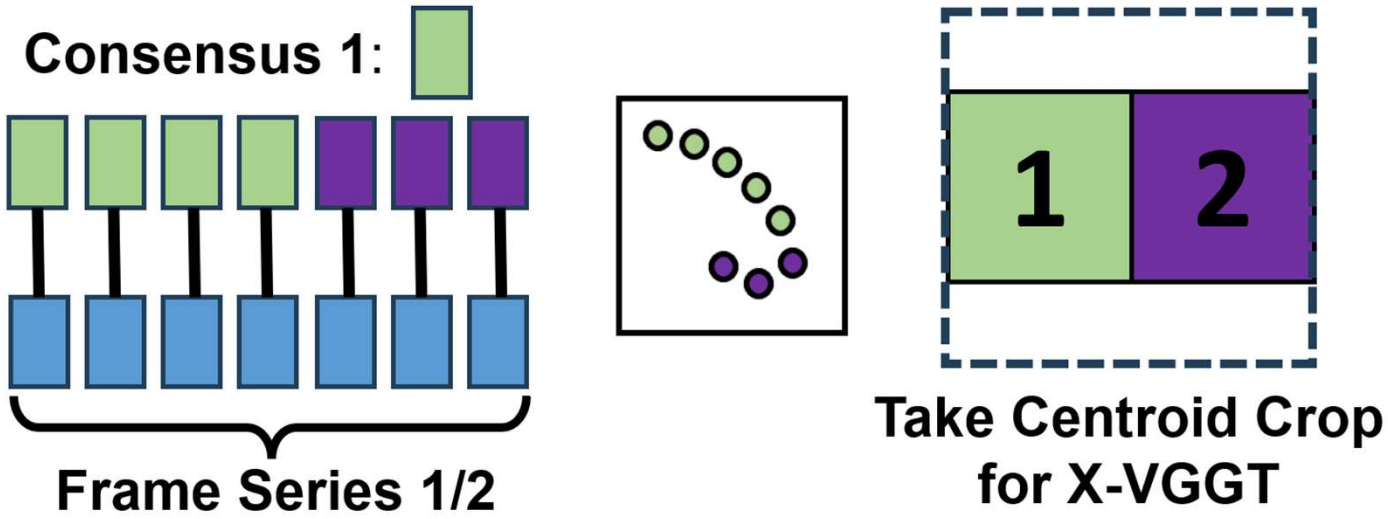
True Query Positions (Street)



GAReT on WRIVA (Handheld)

Adapted from GAReT: Cross-view Video Geolocalization with Adapters and Auto-Regressive Transformers. Credit to Pillai M., et. al.


13 × 24
landscape grid
13
24
interpolate / resize positional grid
GAReT learned positional embeddings
Image Encoders
Frame Series
Patch Gallery
EN-G
EN-A
Nearest Neighbors
Keep Top-1 patch per frame
Frame-to-Patch (F2P)
Frame 1, matched to Reference 1
Frame 1
(Previously Negatives)
Frame 1, matched to Reference 2
Positive
Positive
Positive
Positive
F2P Patch, Ref. 1
F2P Patch, Ref. 2
S2I Patch Ref. 1
S2I Patch Ref. 2





Bursts of
X-VGGT
Predictions

Approach	Score (Mean Error in m)
Baseline GAReT (every frame)	518.57
Baseline GAReT (every 5th frame*) *interpolate in-between positions	481.38
Adapted GAReT (every frame) 384-m center crop	198.93
Adapted GAReT (every 5th frame), 384-m center crop	**195.30 (my best)**
Baseline X-VGGT, 300-m center crop	135.49
Adapted X-VGGT, 300-m center crop	**65.92 (overall best)**
Adapted GAReT + Adapted X-VGGT	**HOPEFULLY TODAY**