Week 5 Progress Report

Generative Satellite Alignment for Cross-View Geolocalization

Student: Wyatt Petula **Mentors**: Soumik Ghosh, Qingyang Liu

Quick Recap of the WRIVA Challenge

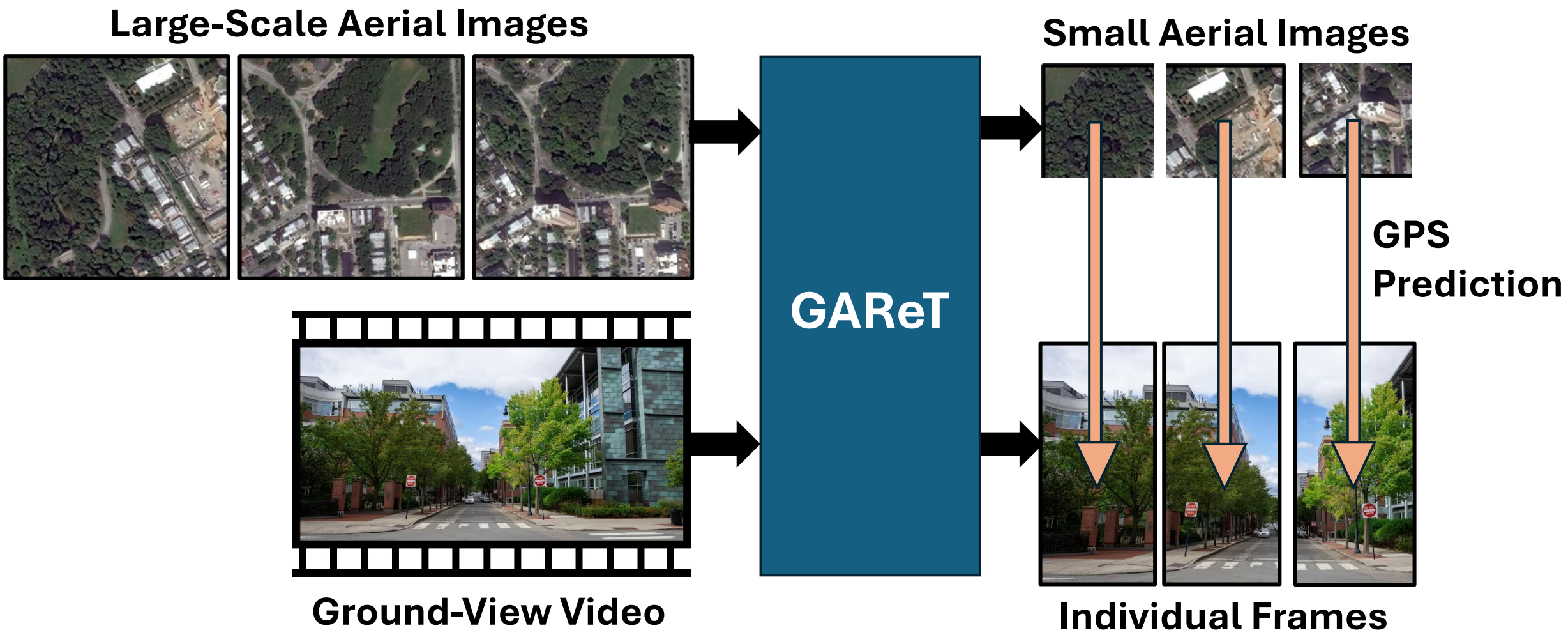
Task:									
#	Participant	Date	ID	avg-geo	001-geo	002-geo	003-geo	004-geo	005
1	SoumikG_UCF	2026-06-19 13:13	804271	**99.02**	184.09	110.5	95.08	142.26	119
2	UCF_Steven	2026-06-17 17:03	801596	**103.87**	190.78	115.18	105.38	147.07	131
3	chandrakanth	2026-06-20 16:59	806005	**107.45**	168.8	115.27	101.85	134.89	99.3
4	weilai	2026-06-05 23:43	782245	**174.26**	284.96	60.74	145.15	151.28	211

Progress

- Studied *GAReT: Cross-view Video Geolocalization with Adapters and Auto-Regressive Transformers* (by Manu S. Pillai, Mamshad N. Rizve, and Mubarak Shah)
- Adapted GAReT to the WRIVA Challenge dataset
- End-to-end baseline inference experiments
- WRIVA Challenge submissions

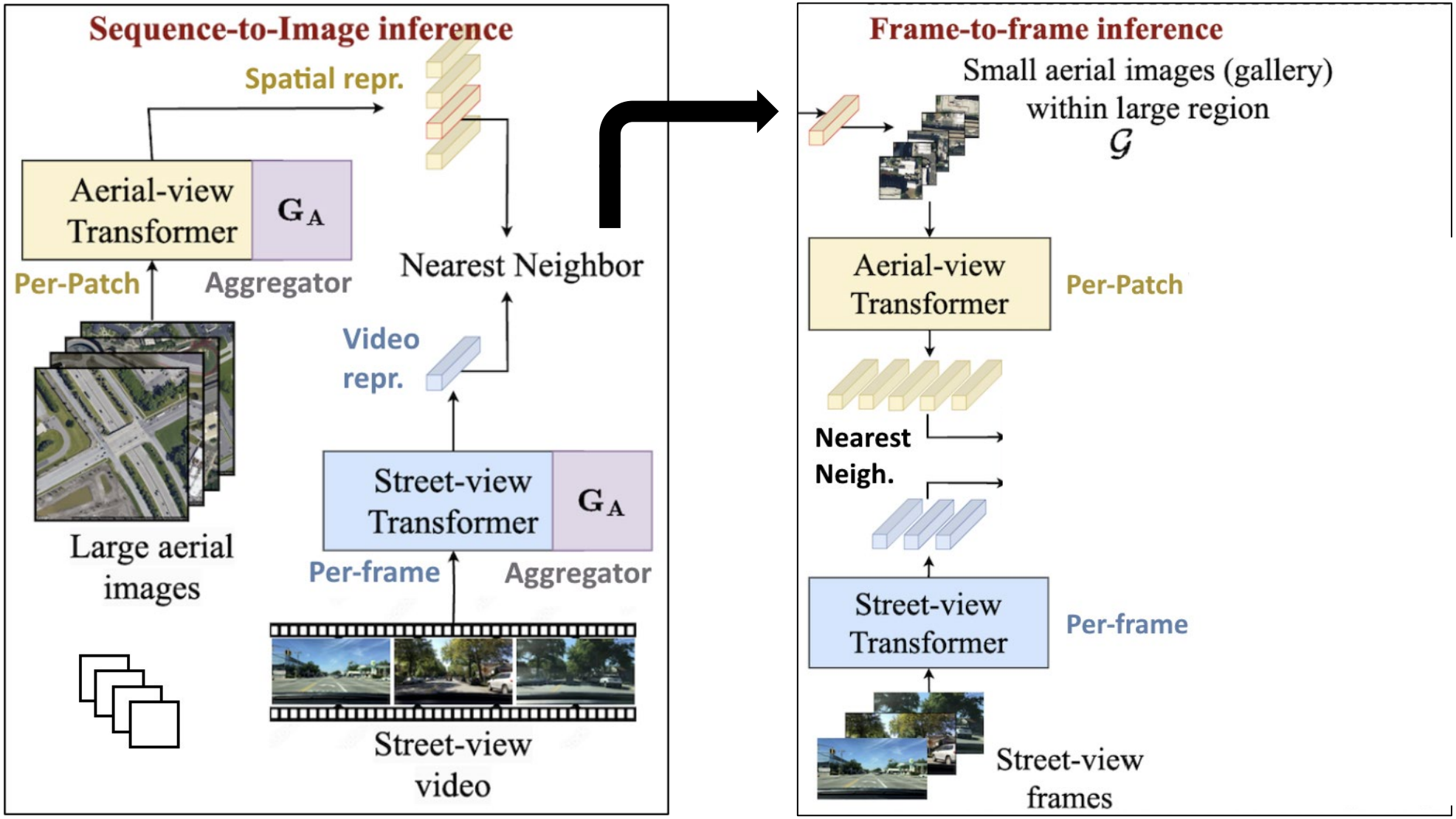
File name	Date	Status	Score
GAReT-Baseline-Codabench-Every5Fram.zip	2026-06-24 18:03	Finished	481.38
SUBMIT_THIS_garet_codabench.zip	2026-06-21 19:24	Finished	518.57

GAReT performs retrieval-based geolocalization



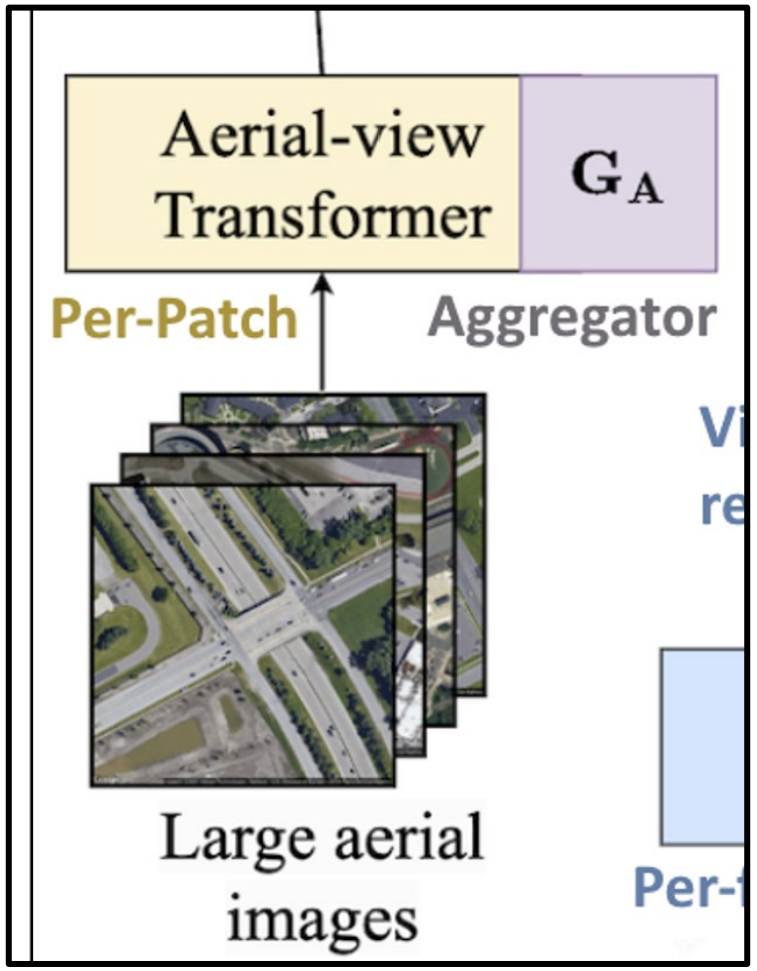
https://www.magnific.com/free-photo/street-with-modern-buildings-green-trees_10352233.htm#fromView=keyword&page=1&position=0&uuid=67e4314b-0e43-4223-92dd-3aab1906185a&query=City+street+with+urban+landscape

GAReT performs retrieval-based geolocalization

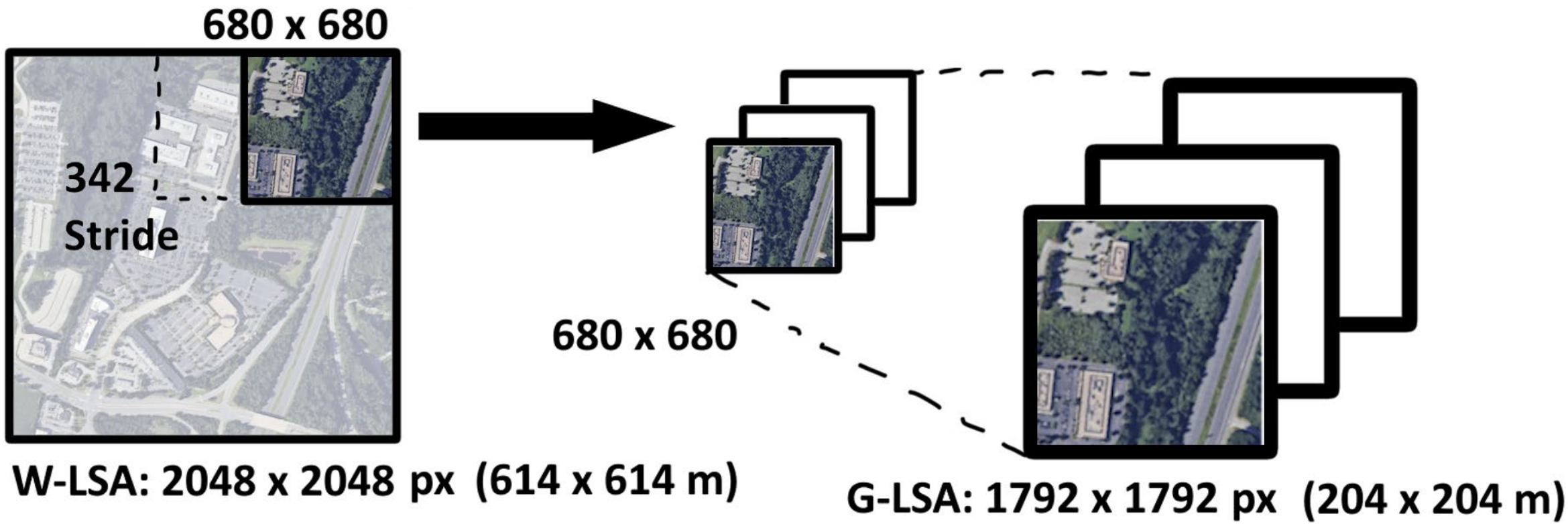


Modified from https://eccv.ecva.net/media/PosterPDFs/ECCV%202024/2461.png?t=1726216876.7051601. Credit to Pillai M., et. al.

GAReT operates at a different spatial resolution (meters/pixel) than the challenge

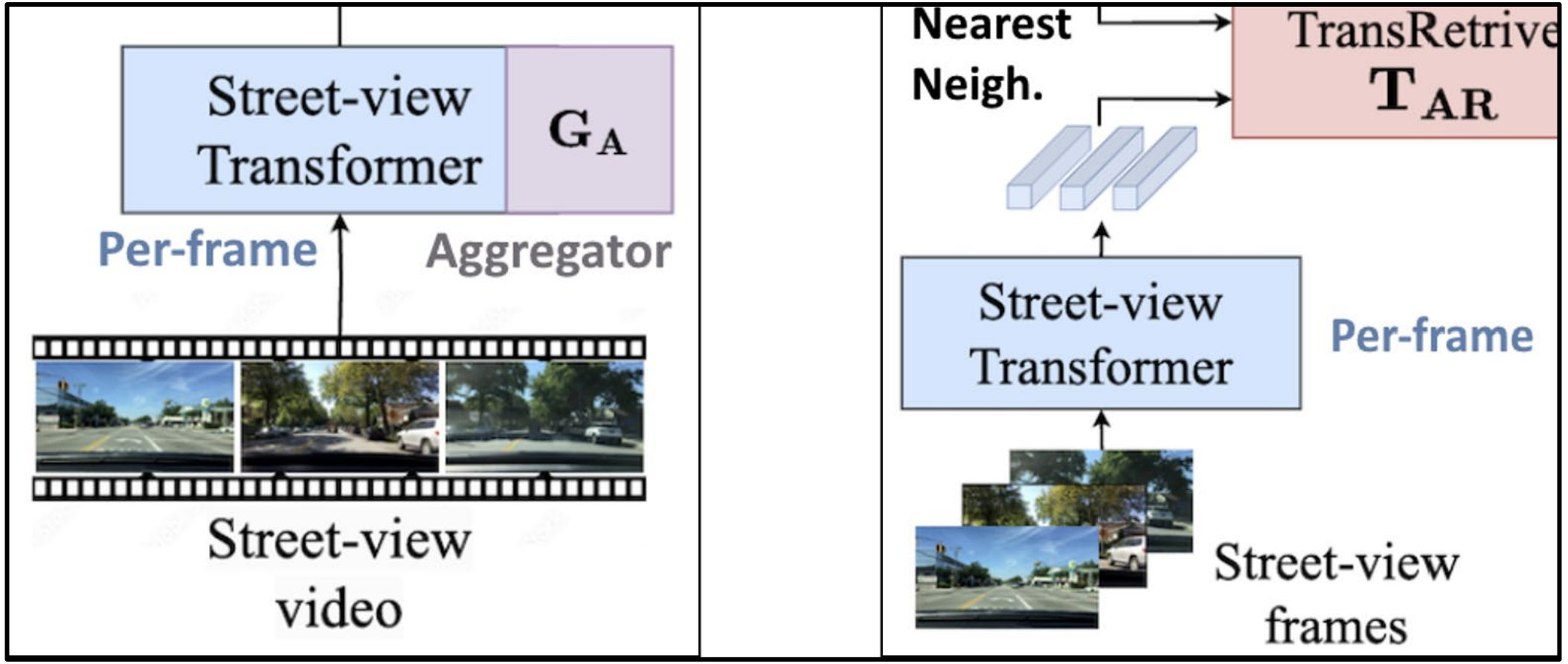


GAReT operates at a different spatial resolution (meters/pixel) than the challenge

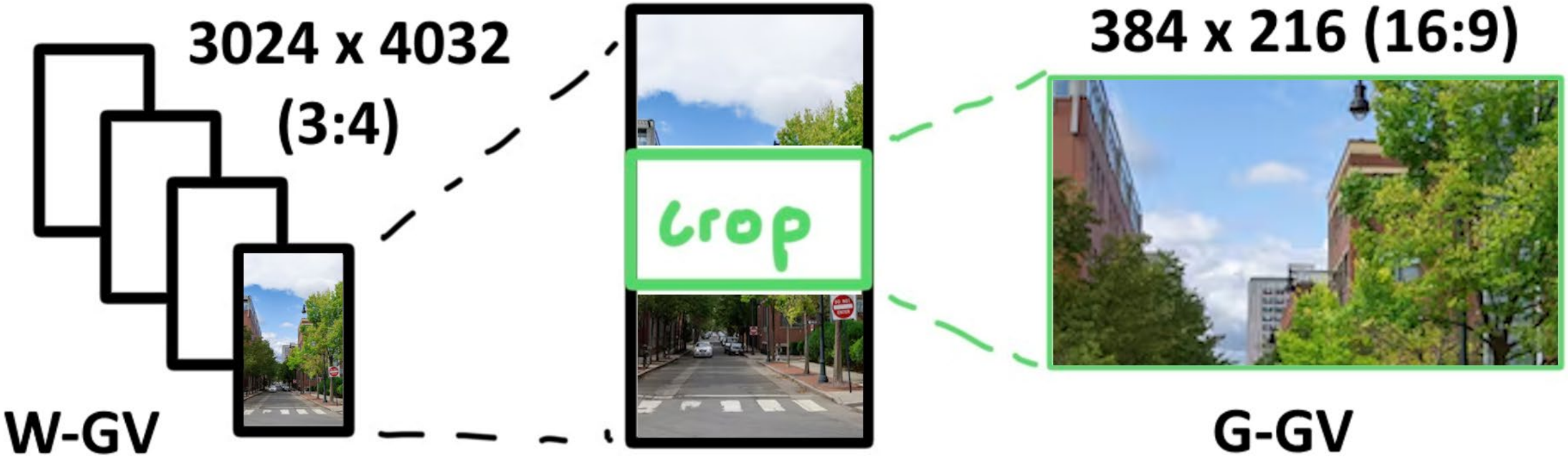


Satellite Image taken from Google Earth

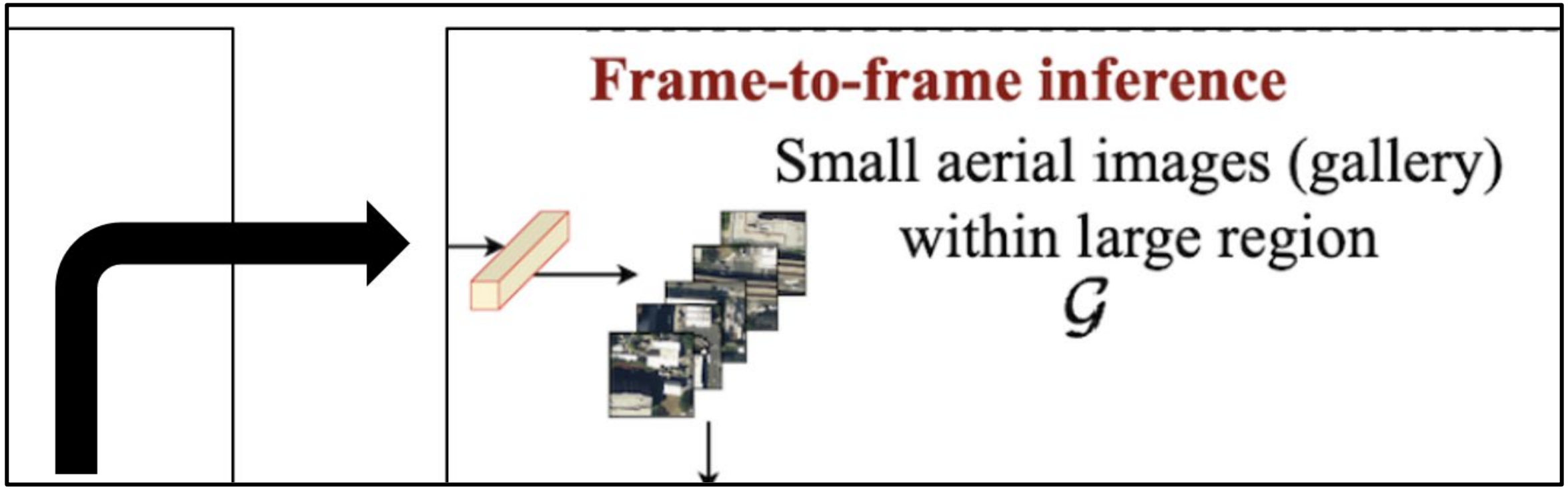
GAReT ground-view images have a different aspect ratio than the challenge



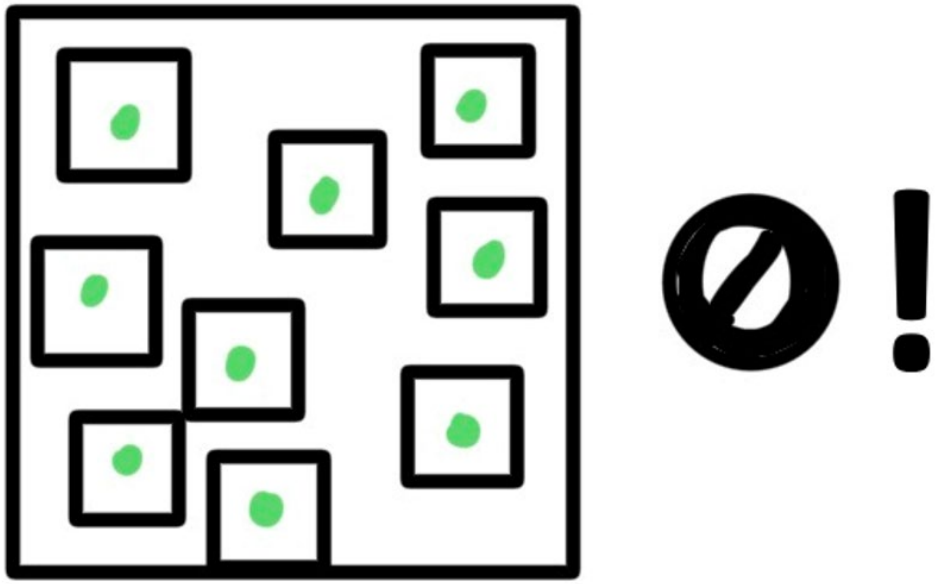
GAReT ground-view images have a different aspect ratio than the challenge



GAReT constructs its small aerial image gallery from the set of ground truth tiles

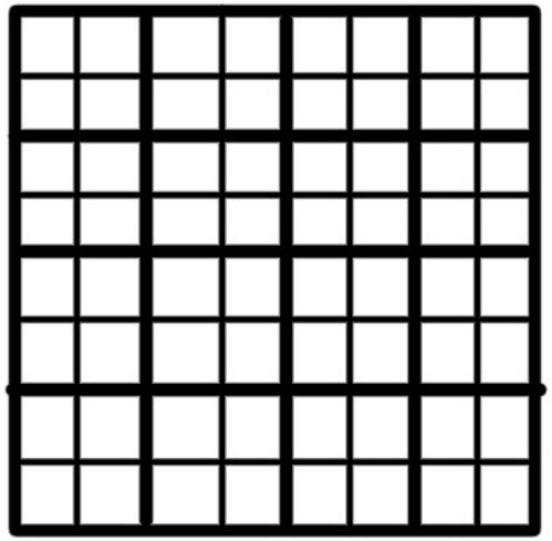


GAReT constructs its small aerial image gallery from the set of ground truth tiles



GAReT: closed set of ground truths

256 x 256 px chips (29 x 29 m)

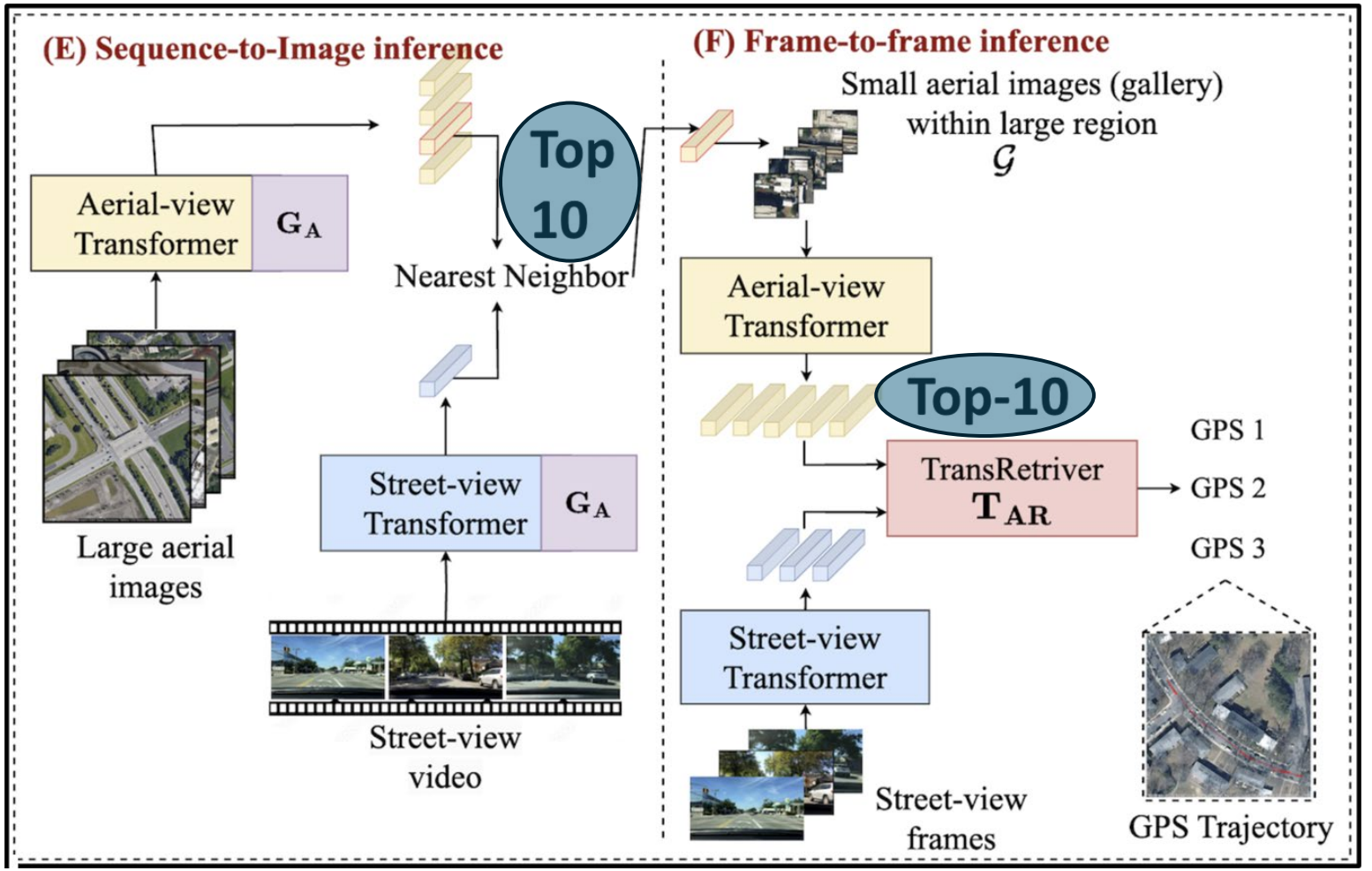


128 px str.

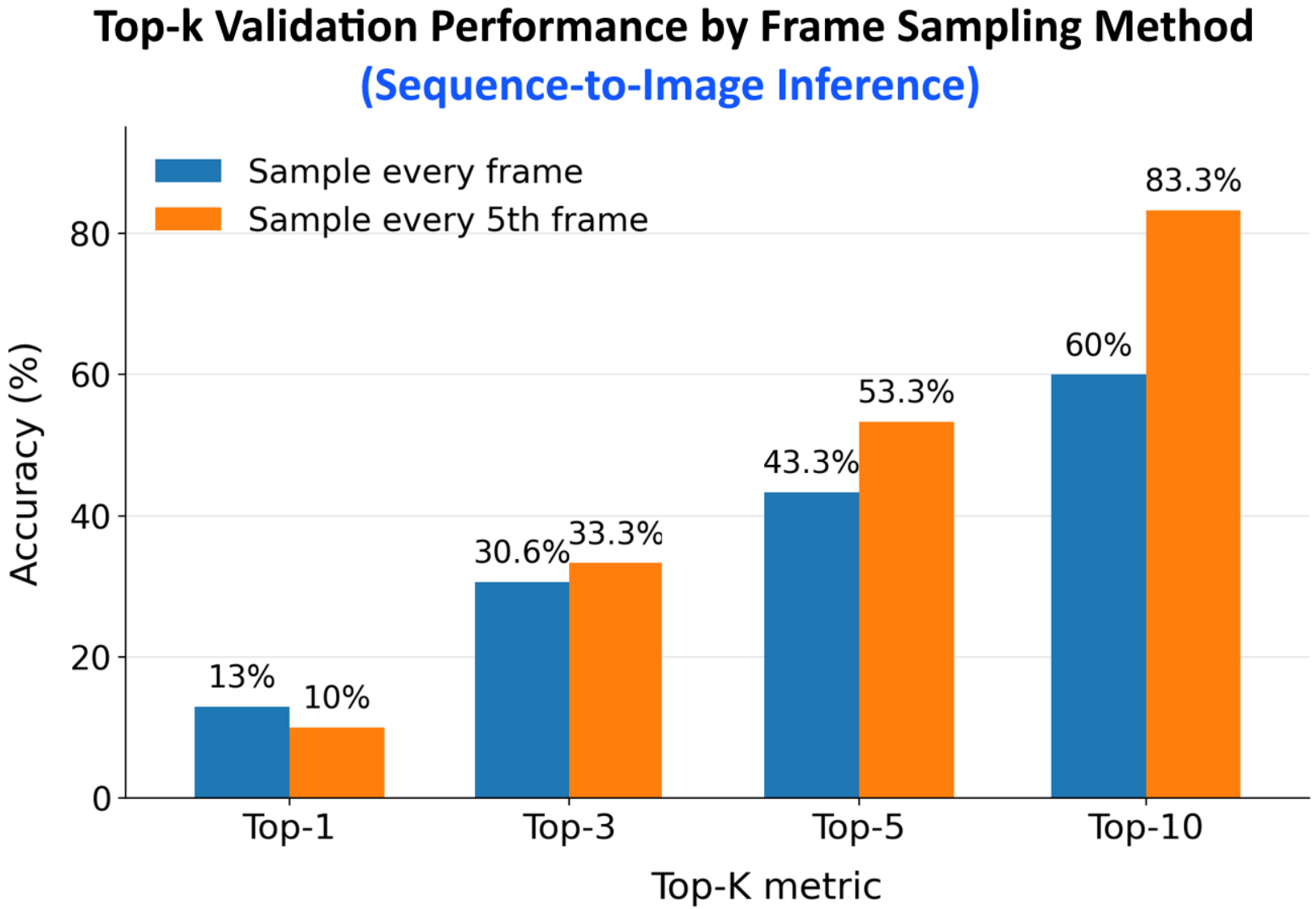
WRIVA Adaptation: patchify LSAs

I did two end-to-end inference experiments

(1) Sequence from all frames **(2)** Use Every 5th Frame

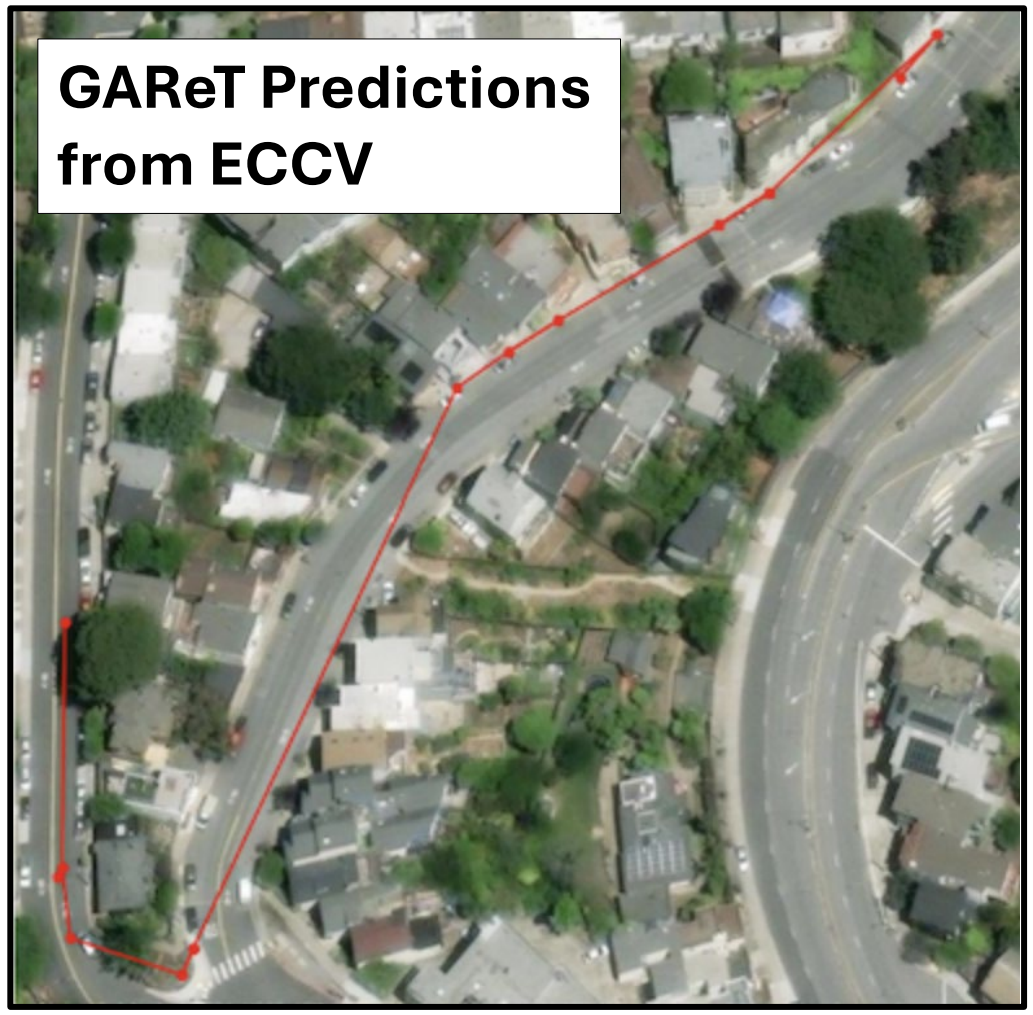


I did two end-to-end inference experiments



Score	TEST SET
481.38	Every 5th Frame
518.57	Every Frame

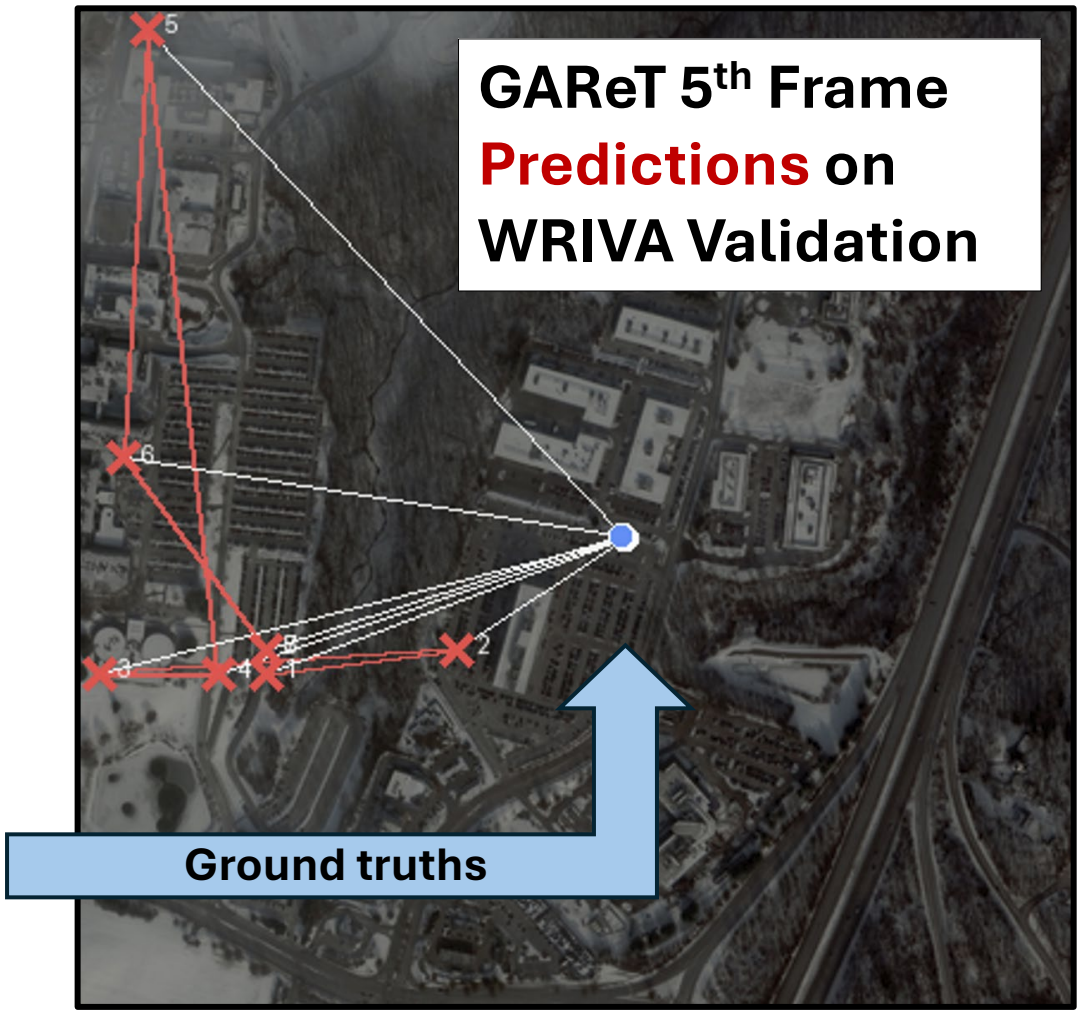
The inference results can be explained by data mismatch







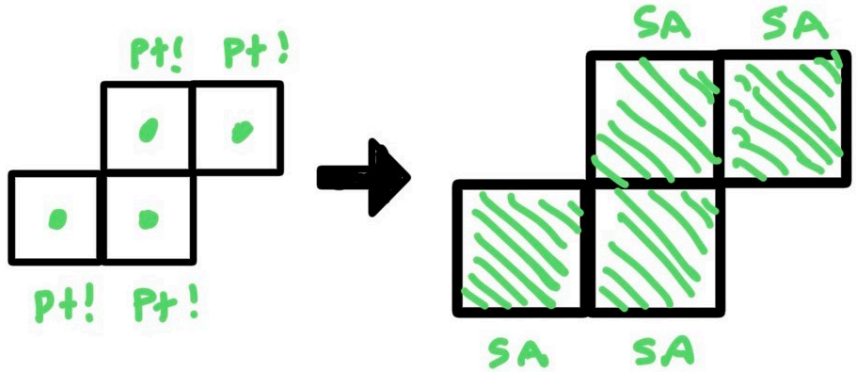




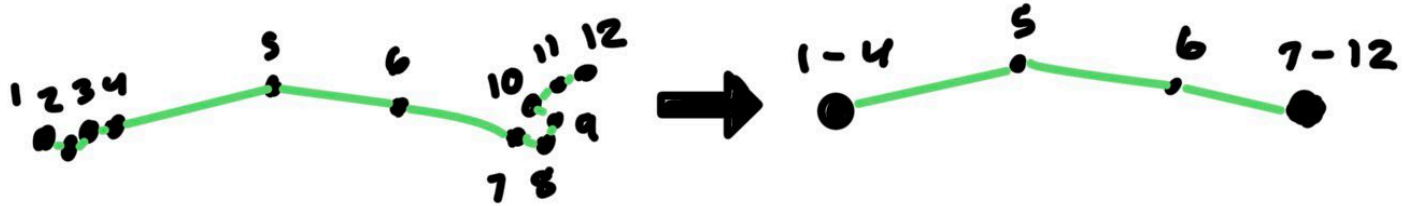
Modified from https://eccv.ecva.net/media/PosterPDFs/ECCV%202024/2461.png?t=1726216876.7051601. Credit to Pillai M., et. al.

The plan moving forward is to address the retrieval bottleneck

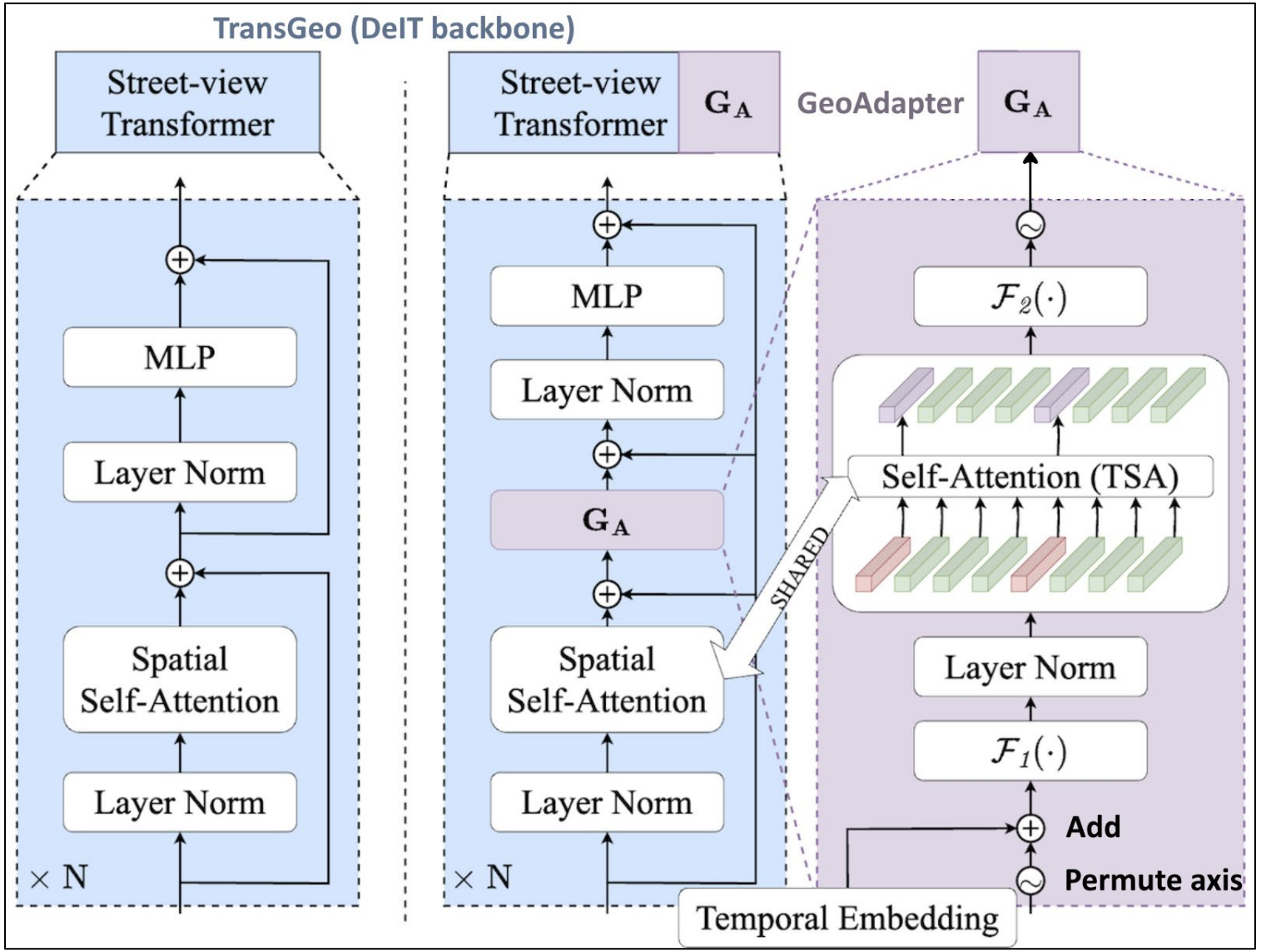
- **Priority:** Use GAReT's end-to-end for retrieval



- **Supplementary:** Fine-tune GAReT on the VisymScenes dataset
- **Small Idea:** Treat clumped WRIVA ground views as a single point

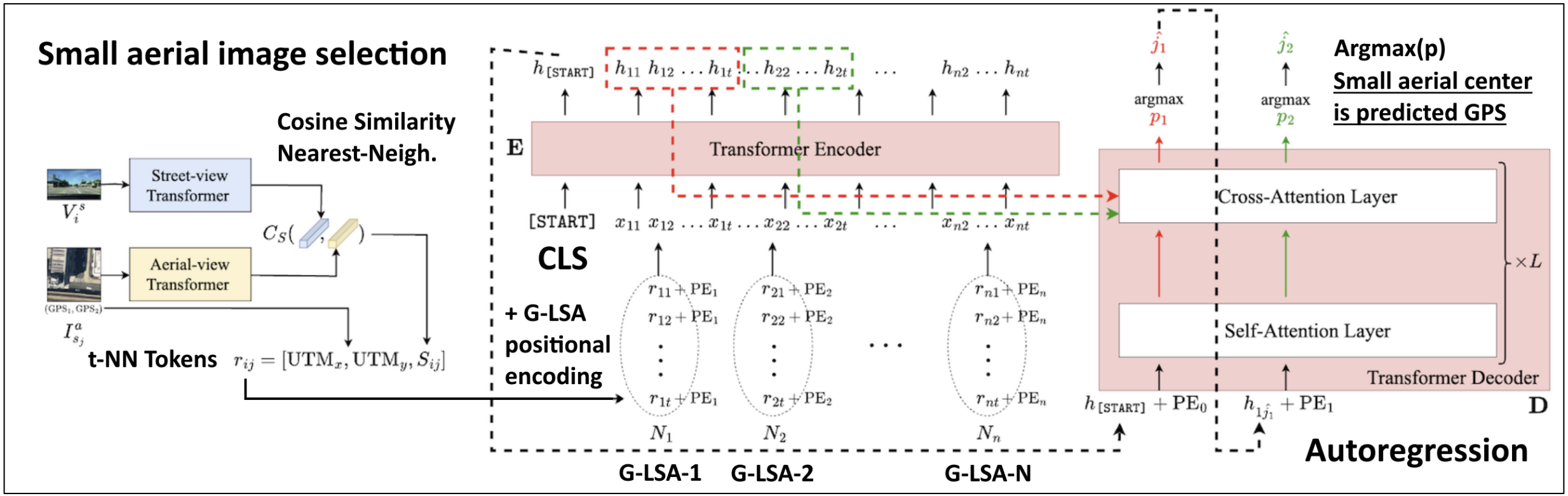


Transformer-GeoAdapter architecture



Modified from https://eccv.ecva.net/media/PosterPDFs/ECCV%202024/2461.png?t=1726216876.7051601. Credit to Pillai M., et. al.

Autoregressive TransRetriever architecture



Modified from https://eccv.ecva.net/media/PosterPDFs/ECCV%202024/2461.png?t=1726216876.7051601. Credit to Pillai M., et. al.