Week 6 Progress Report

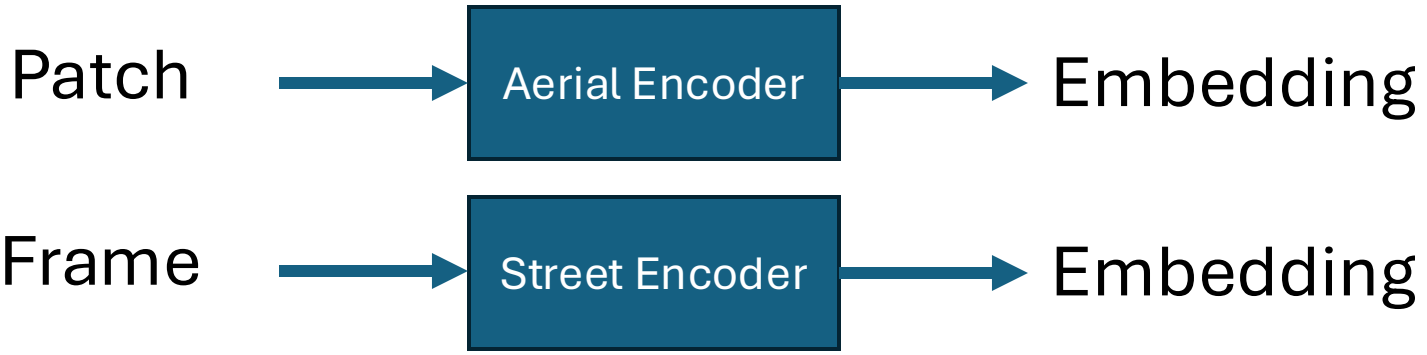
Generative Satellite Alignment for Cross-View Geolocalization

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

-- Adapting GAReT to the WRIVA Challenge --

Fine-Tuning GAReT on the WRIVA challenge

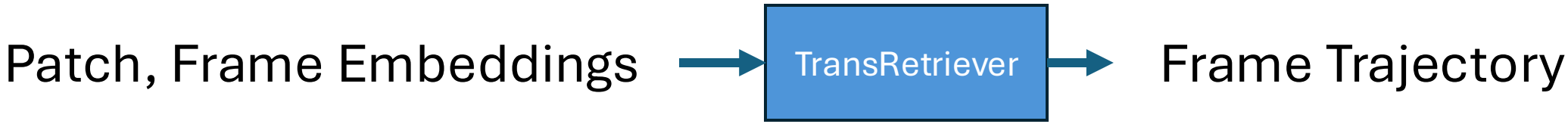
#1: Street and Aerial Image Encoders ✓



#2: GeoAdapter X



#3: TransRetriever X



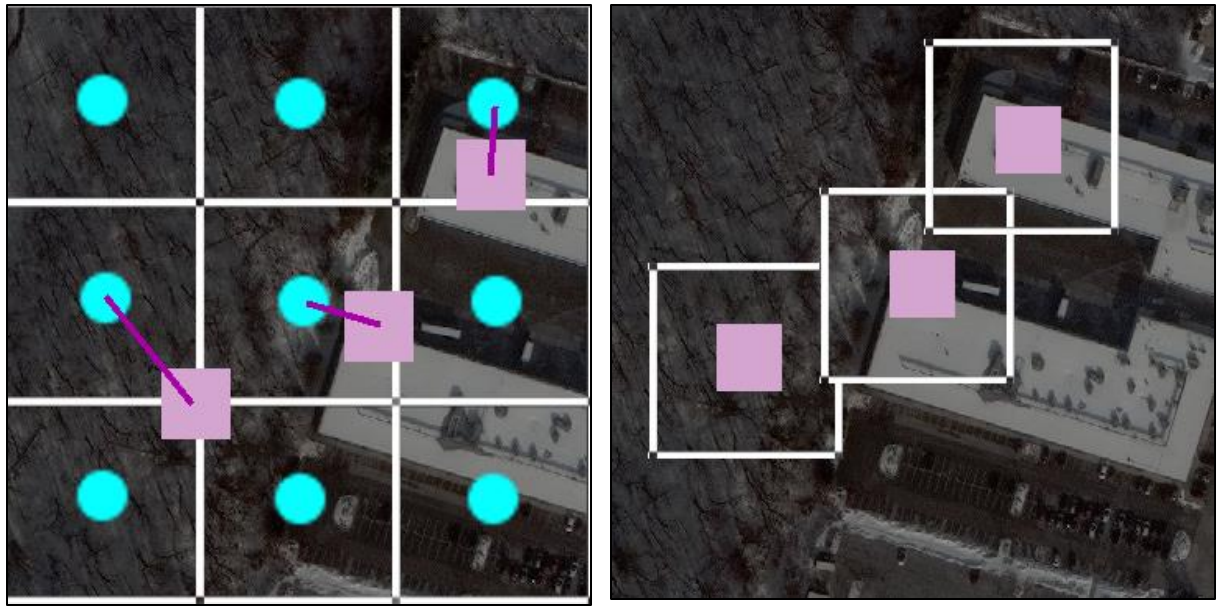
Reformatting the challenge dataset as GAMa

GAMa Dataset Setup

1 Large Aerial / Sequence



Small Aerial Gallery



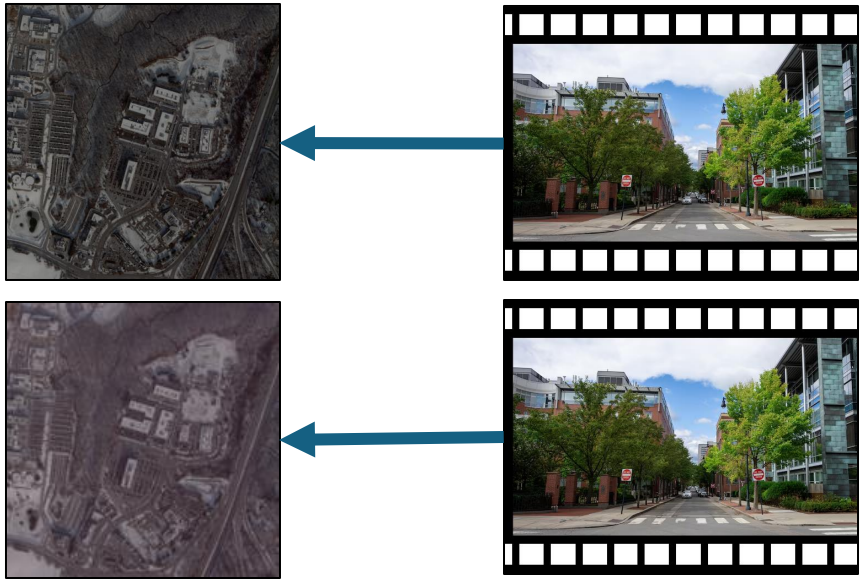
Uncentered

Centered

Reformatting the challenge dataset as GAMa

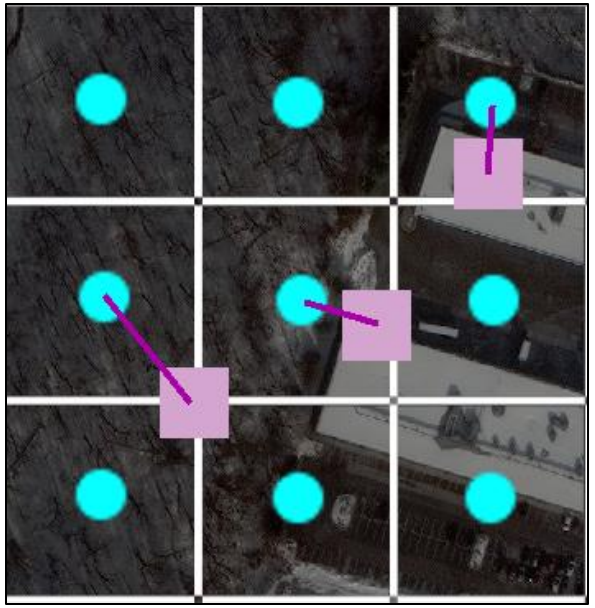
Adapted Challenge Dataset Setup

Multiple Large Aerials / Sequence



... ...

Small Aerial Gallery



Must use uncentered

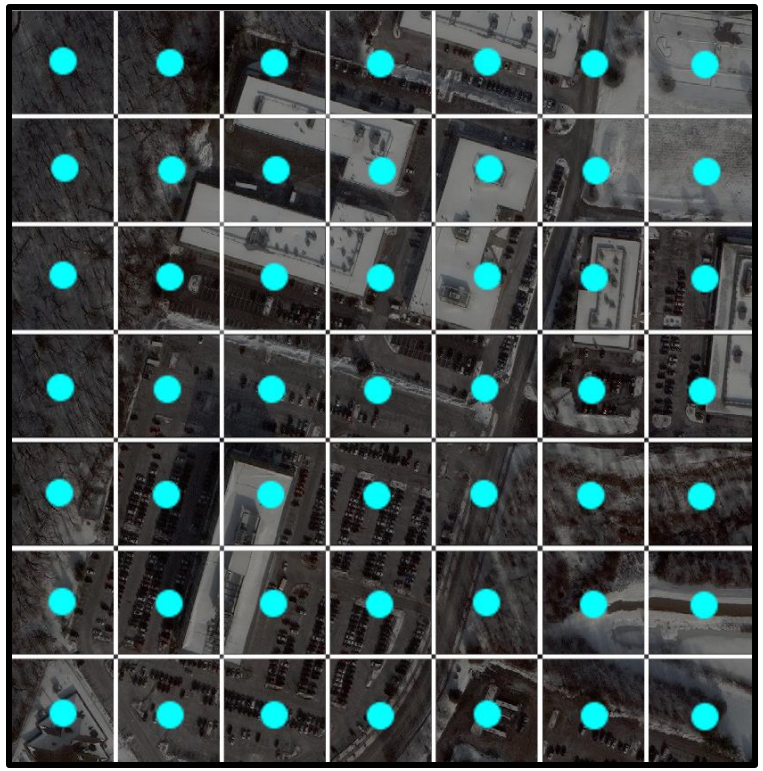
Adapting GAReT to the challenge

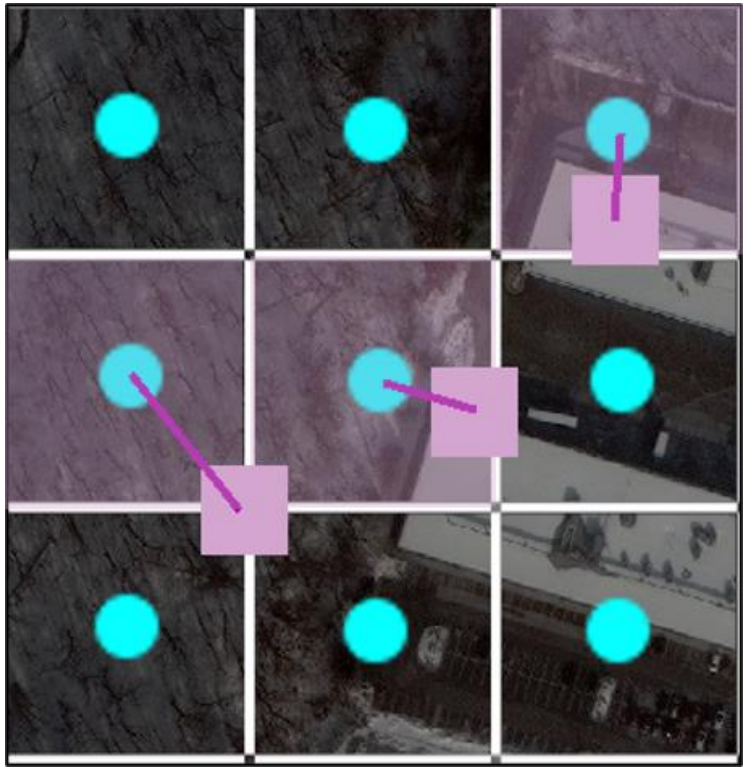
Minimize GPS Approximation Error

7x7 patches, 256 x 256 pixels, 0% overlap

30-meter max error

Garet





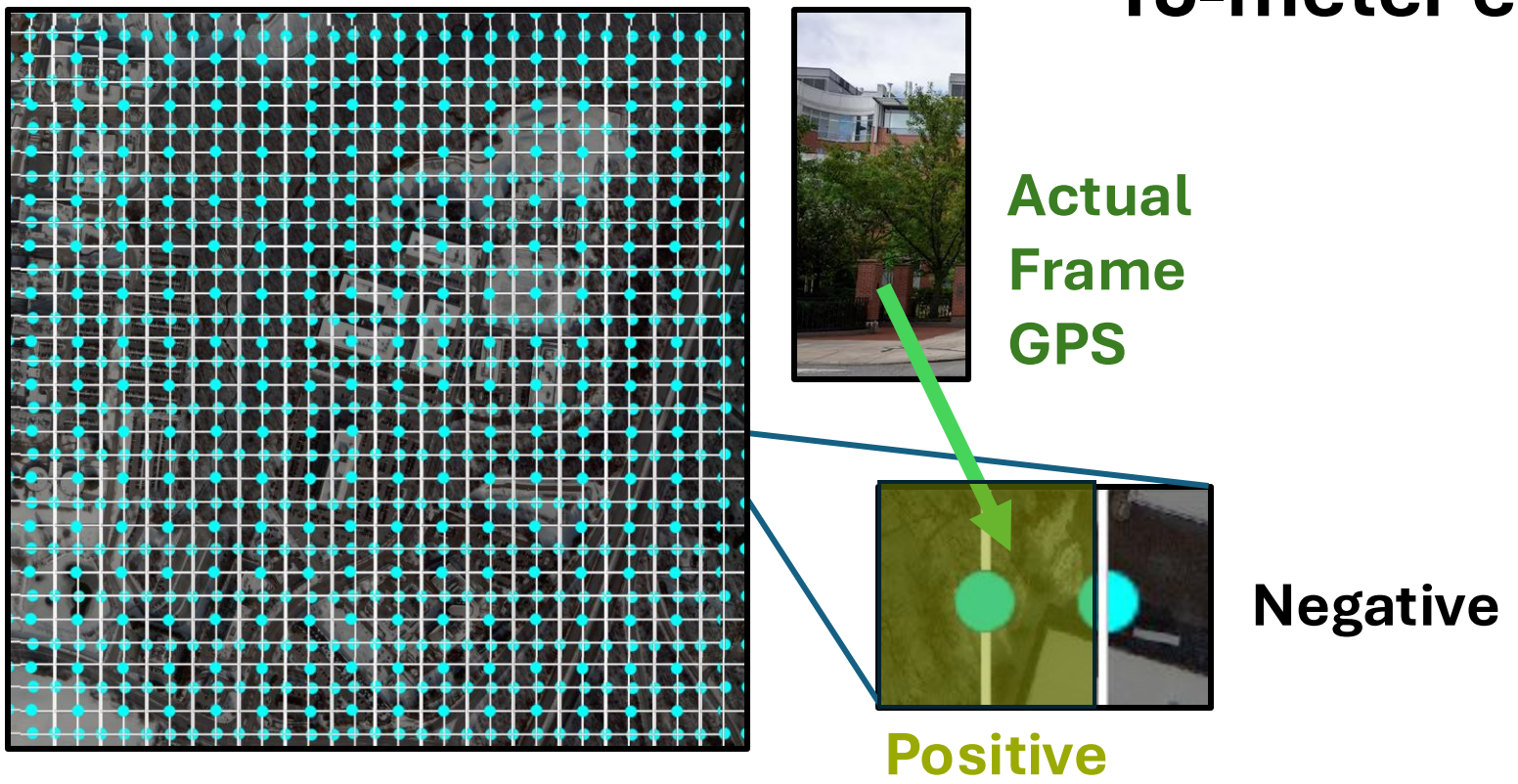
Adapting GAReT to the challenge

Minimize GPS Approximation Error

16x16 patches, 128 x 128 pixels, 50% overlap

15-meter error

Challenge Adaptation



Training and validation set preparation

Removed Data

- Corrupted large aerial images
- Bad sites listed by challenge organizers
- Sites with out-of-bounds ground GPS

Balancing Representation: Large Aerials / Site

- Training set: sample four
  - Sites: 106 Min: 1 Median: 7 Max: 114
- Validation set: sample six
  - Sites: 3 Min: 4 Median: 6 Max: 26

Training and validation set preparation

Final Dataset Metrics

Training Set:

- Sites: 115
- Raw ground-small aerial pairs: 2.35M
- Balanced pairs used: 476K
- Unique small aerial references: 6,548

Validation Set:

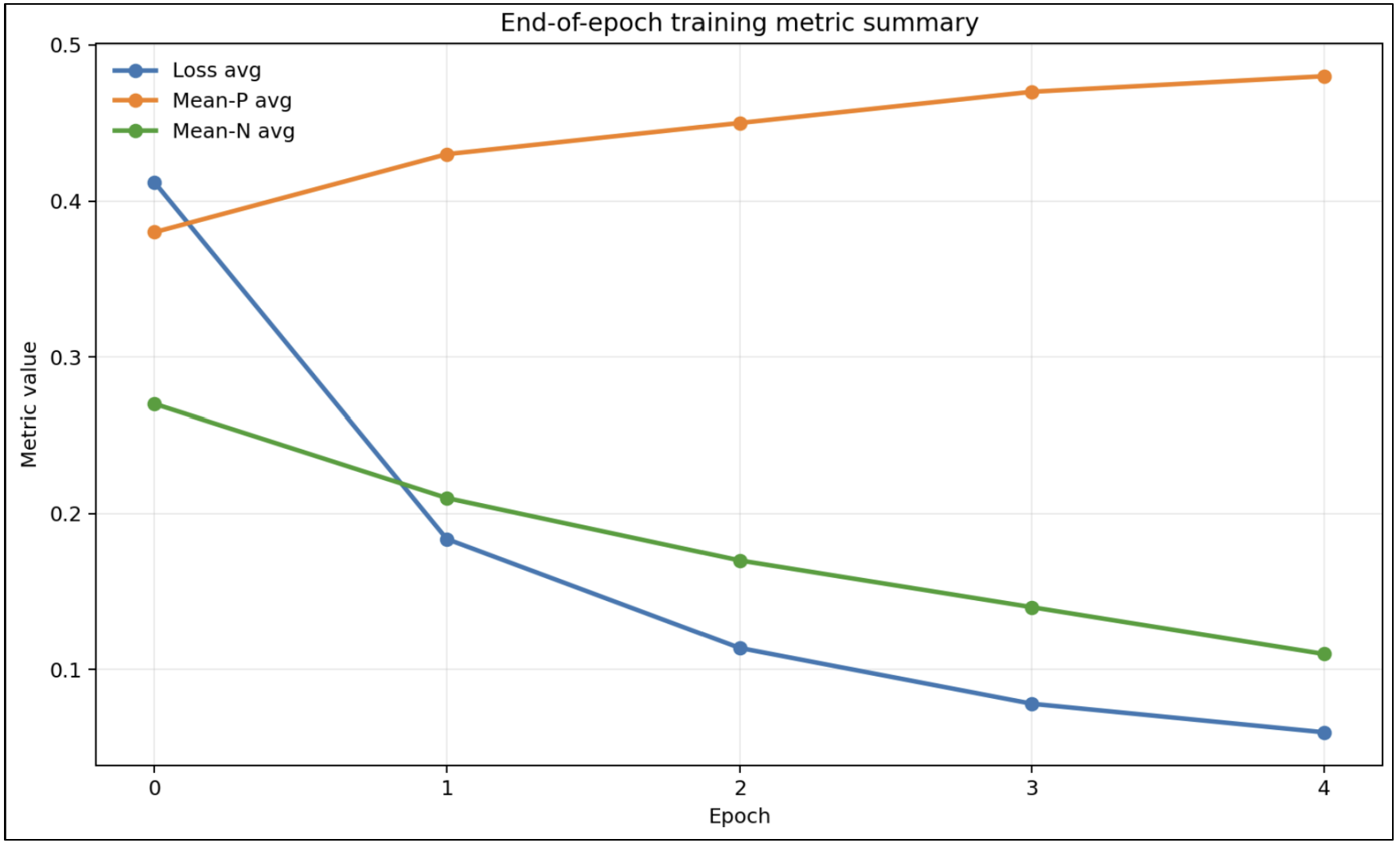
- Sites: 3
- Raw ground-small aerial pairs: 14.4K
- Balanced pairs used: 6.4K
- Unique small aerial references: 100

A conservative fine-tuning of the encoders shows GAReT is learning the training set

Average Loss

Average Positive Similarity/Score

Average Negative Similarity/Score



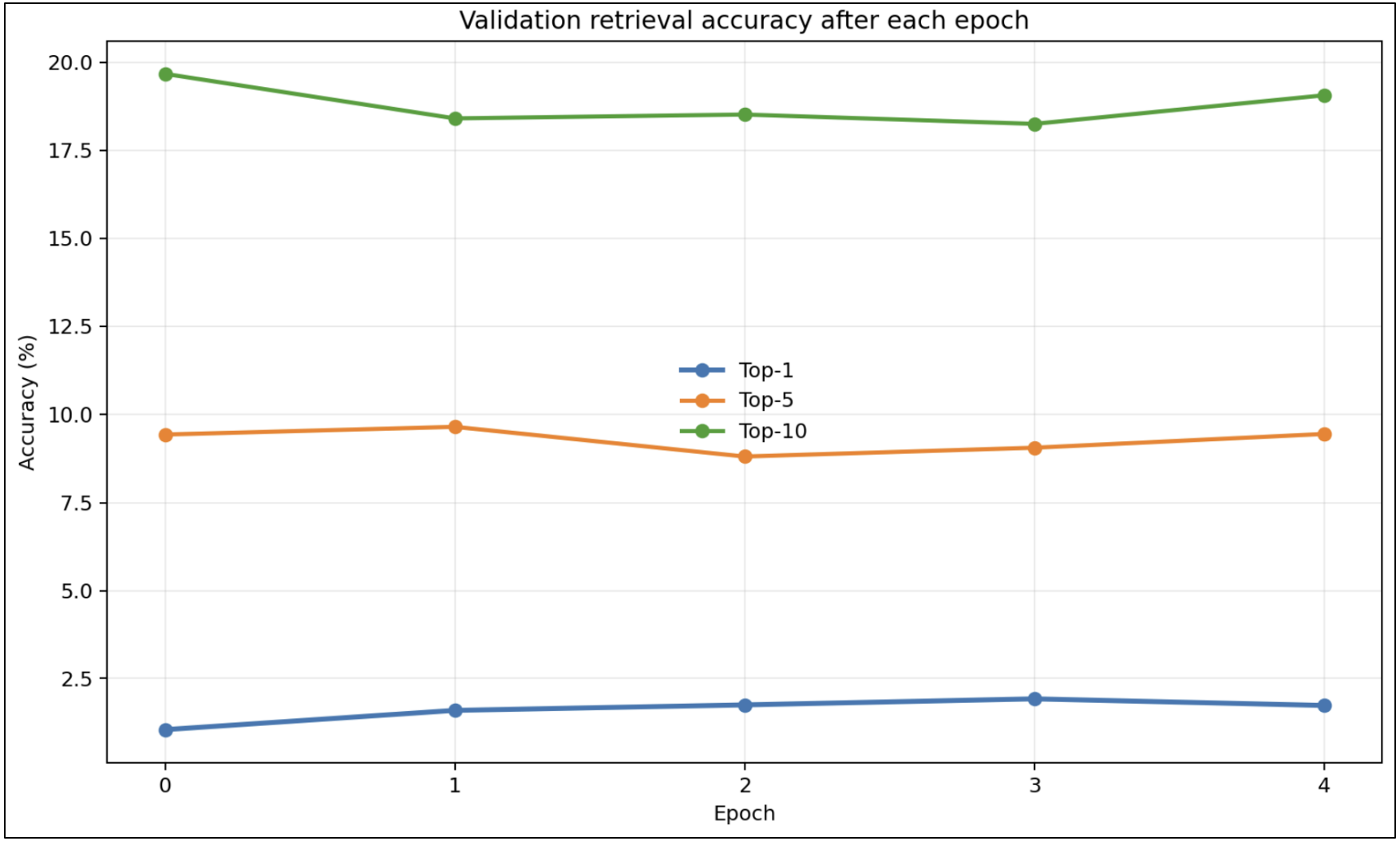
- **Epochs**: 5
- **Batch Size**: 64
- **Learning Rate**: 1e-05
- **Optimizer**: adamw

The fine-tuning shows GAReT not generalizing to the validation set

Top-10 Accuracy

Top-5 Accuracy

Top-1 Accuracy



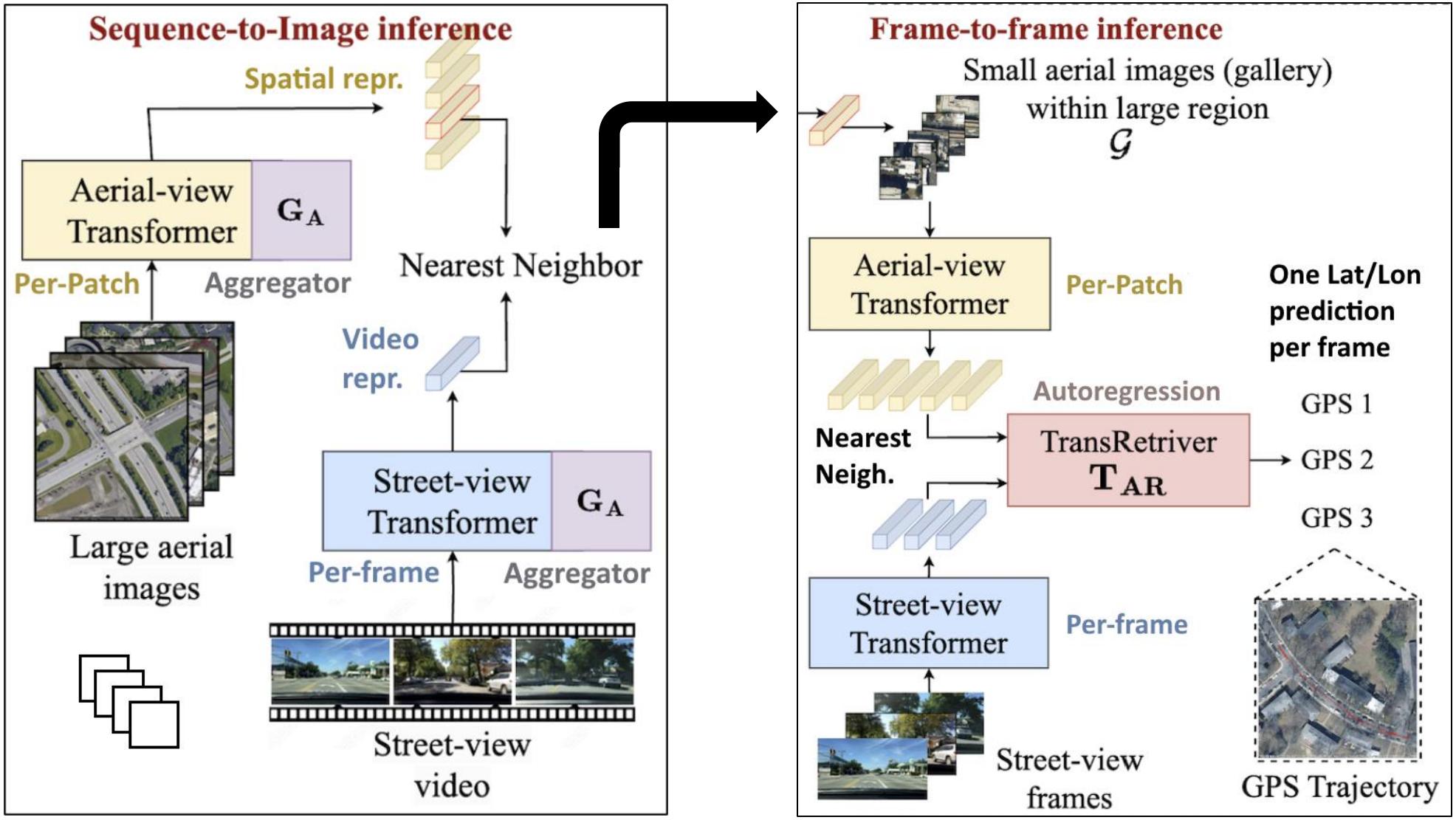
- **Epochs**: 5
- **Batch Size**: 64
- **Learning Rate**: 1e-05
- **Optimizer**: adamw

Submission results are marginally better than previously

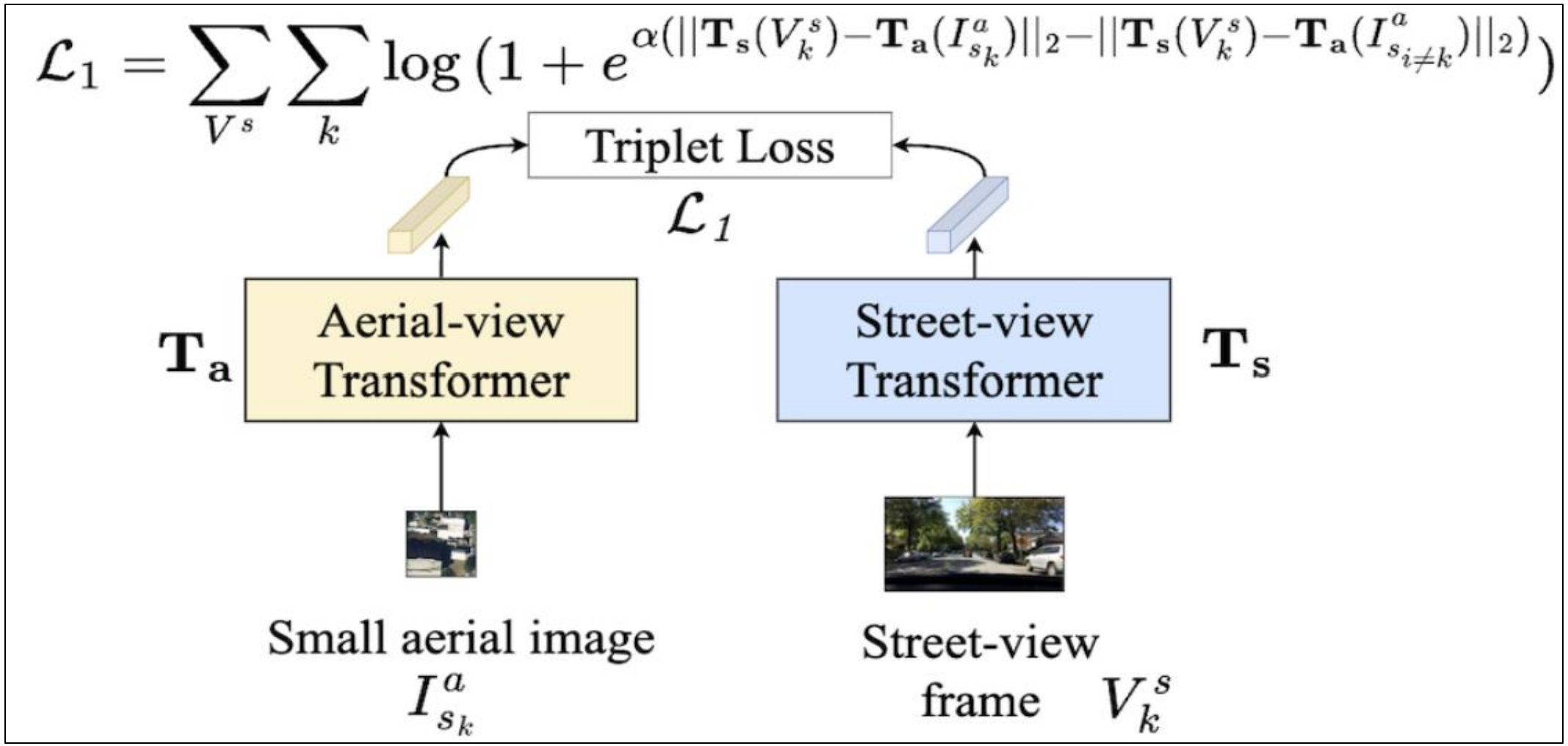
File name	Date	Status	Score
garet_test-ft_encoders_codabench.zip	2026-07-01 23:26	Finished	480.07
GAReT-Baseline-Codabench-Every5Fram.zip	2026-06-24 18:03	Finished	481.38
SUBMIT_THIS_garet_codabench.zip	2026-06-21 19:24	Finished	518.57

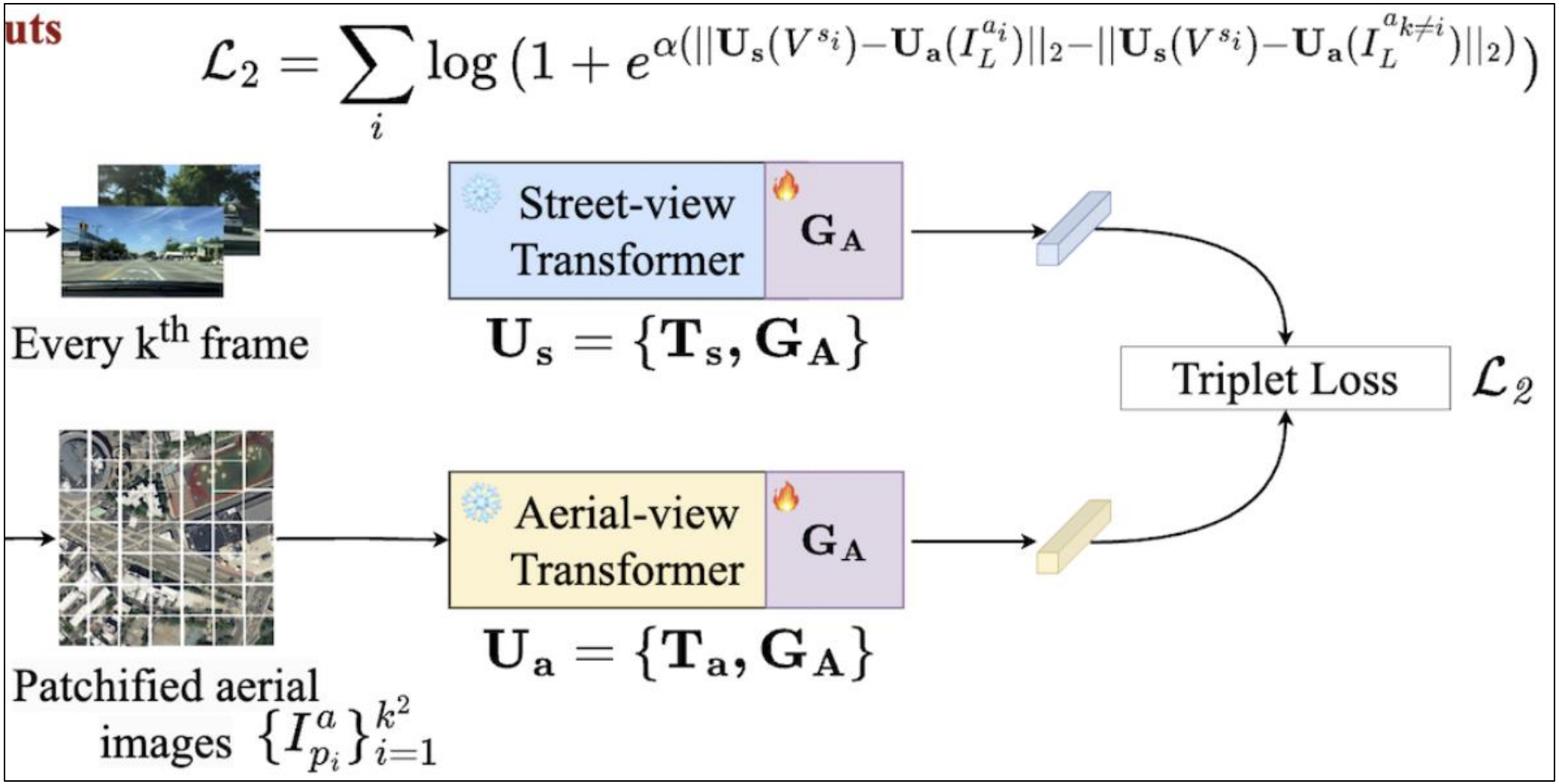
Goals for next week

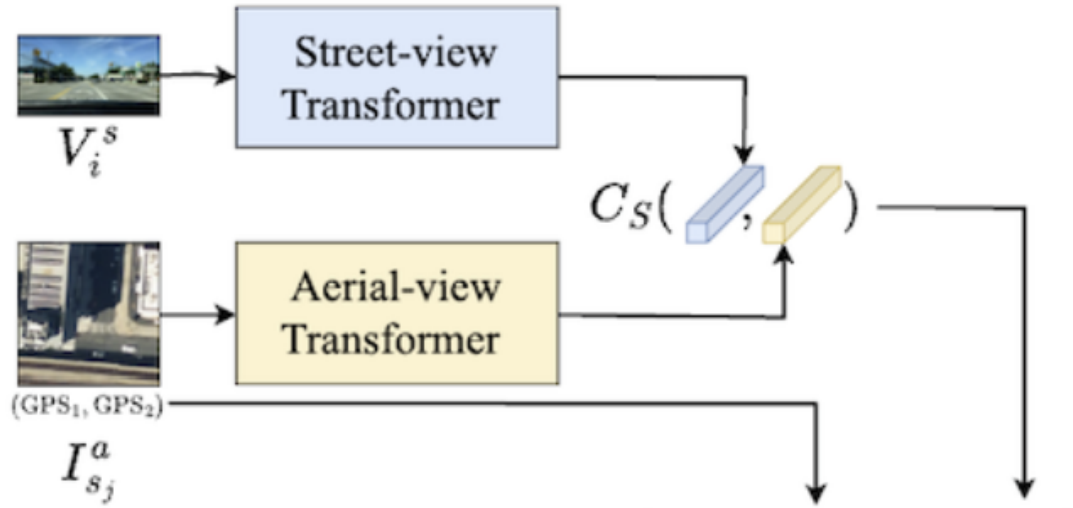
- Improve aerial and street image encoder fine-tuning
- Fine-tune ~~GeoAdapter and~~ TransRetriever
- Remove (and possibly rework) Sequence-to-Image inference


Sequence-to-Image inference
Spatial repr.
Aerial-view Transformer
G_A
Per-Patch
Aggregator
Nearest Neighbor
Video repr.
Large aerial images
Street-view Transformer
G_A
Per-frame
Aggregator
Street-view video
Frame-to-frame inference
Small aerial images (gallery) within large region
$\mathcal{G}$
Aerial-view Transformer
Per-Patch
One Lat/Lon prediction per frame
Autoregression
Nearest Neigh.
TransRetriver
T_{AR}
GPS 1
GPS 2
GPS 3
Street-view Transformer
Per-frame
Street-view frames
GPS Trajectory

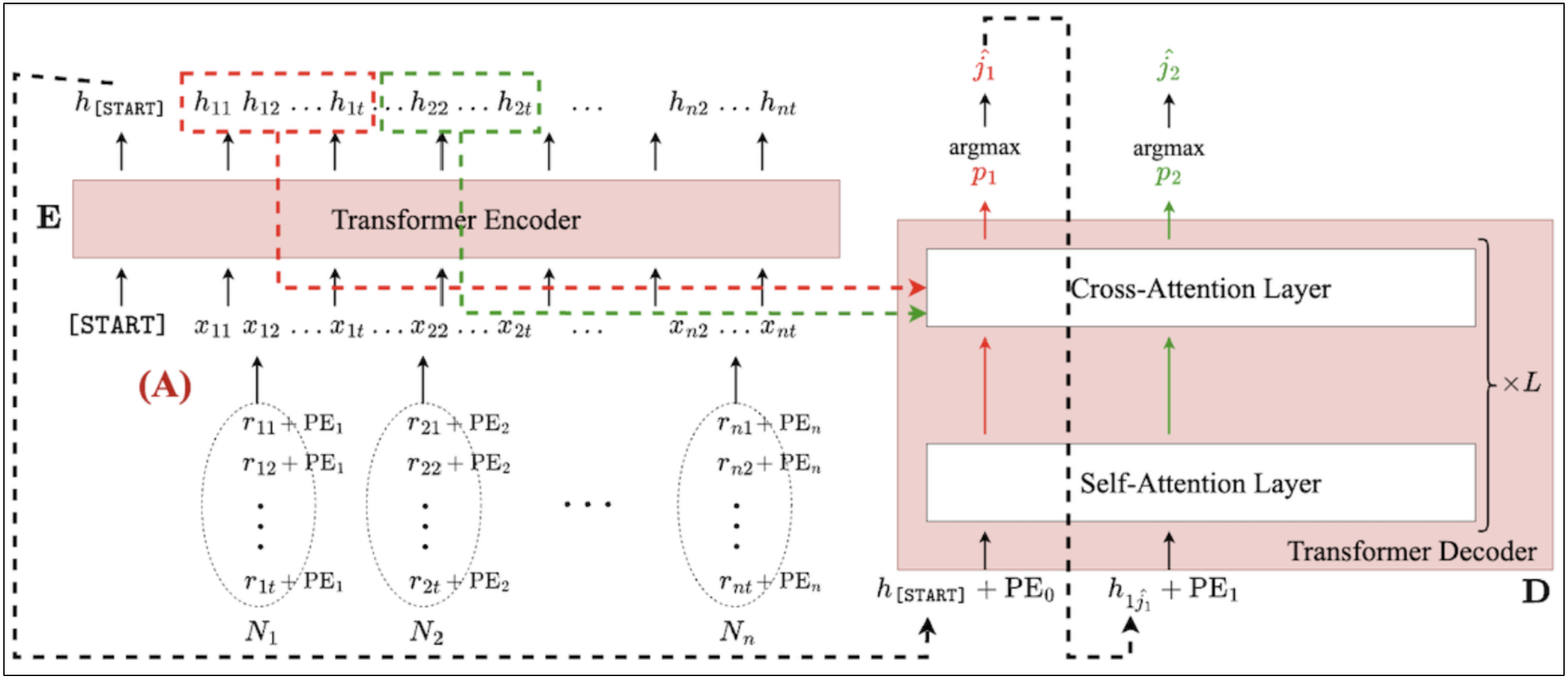
$$\mathcal{L}_1 = \sum_{V^s} \sum_{k} \log\left(1 + e^{\alpha(||\mathbf{T_s}(V_k^s) - \mathbf{T_a}(I_{s_k}^a)||_2 - ||\mathbf{T_s}(V_k^s) - \mathbf{T_a}(I_{s_{i\neq k}}^a)||_2)}\right)$$




uts
$\mathcal{L}_2 = \sum_i \log\left(1 + e^{\alpha(||\mathbf{U_s}(V^{s_i}) - \mathbf{U_a}(I_L^{a_i})||_2 - ||\mathbf{U_s}(V^{s_i}) - \mathbf{U_a}(I_L^{a_{k\neq i}})||_2)}\right)$
Every k^th frame
Street-view Transformer
$\mathbf{G_A}$
$\mathbf{U_s} = \{\mathbf{T_s}, \mathbf{G_A}\}$
Triplet Loss
$\mathcal{L}_2$
Patchified aerial images $\{I_{p_i}^a\}_{i=1}^{k^2}$
Aerial-view Transformer
$\mathbf{G_A}$
$\mathbf{U_a} = \{\mathbf{T_a}, \mathbf{G_A}\}$



$$\mathcal{L}_3 = -\sum_i \sum_t y_t \log\left(1 - p_t\right)$$


$h_{[\text{START}]}$
h_{11} h_{12} ... h_{1t} ... h_{22} ... h_{2t} ... h_{n2} ... h_{nt}
E
Transformer Encoder
[START]
x_{11} x_{12} ... x_{1t} ... x_{22} ... x_{2t} ... x_{n2} ... x_{nt}
(A)
r_{11} + PE_1
r_{12} + PE_1
r_{1t} + PE_1
r_{21} + PE_2
r_{22} + PE_2
r_{2t} + PE_2
r_{n1} + PE_n
r_{n2} + PE_n
r_{nt} + PE_n
N_1
N_2
N_n
$\hat{j}_1$
$\hat{j}_2$
argmax p_1
argmax p_2
Cross-Attention Layer
Self-Attention Layer
$\times L$
Transformer Decoder
$h_{[\text{START}]}$ + PE_0
$h_{1\hat{j}_1}$ + PE_1
D