Week 8 Progress Report

Generative Satellite Alignment for Cross-View Geolocalization

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

-- Fine-Tuning GAReT and 3D Opportunities --

Terminology moving forward



Reference Image

WRIVA Challenge

S2I Patch

Sequence-to-Image,
“Large Aerial Images”

F2F Patch

Frame-to-Frame,
“Small Aerial Images”

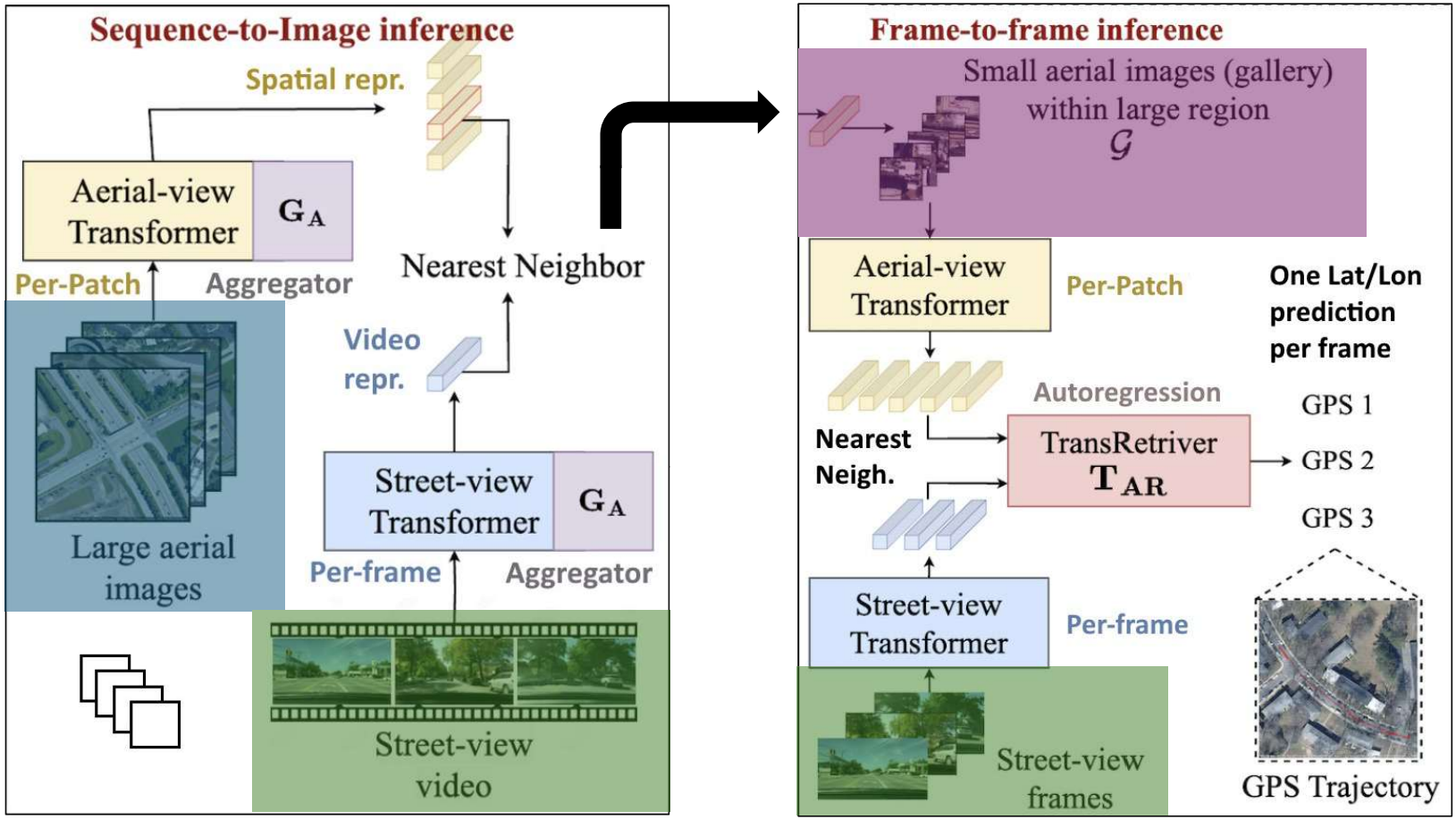
https://www.magnific.com/free-photo/street-with-modern-buildings-green-trees_10352233.htm#fromView=keyword&page=1&position=0&uuid=67e4314b-0e43-4223-92dd-3aab1906185a&query=City+street+with+urban+landscape

Query

Frames and
Sequences



Terminology applied to GAReT's architecture



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

Differences between GAReT and WRIVA

WRIVA_Challenge_Dataset

--Site_001

----sequential_query_frames



Pixel Resolution: 3024 x 4032
(Portrait)

----reference_images



Pixel Resolution: 2048 x 2048

--Site_002

-- ...

GAReT (using GAMa)



Pixel Resolution:
384 x 216 (Landscape)

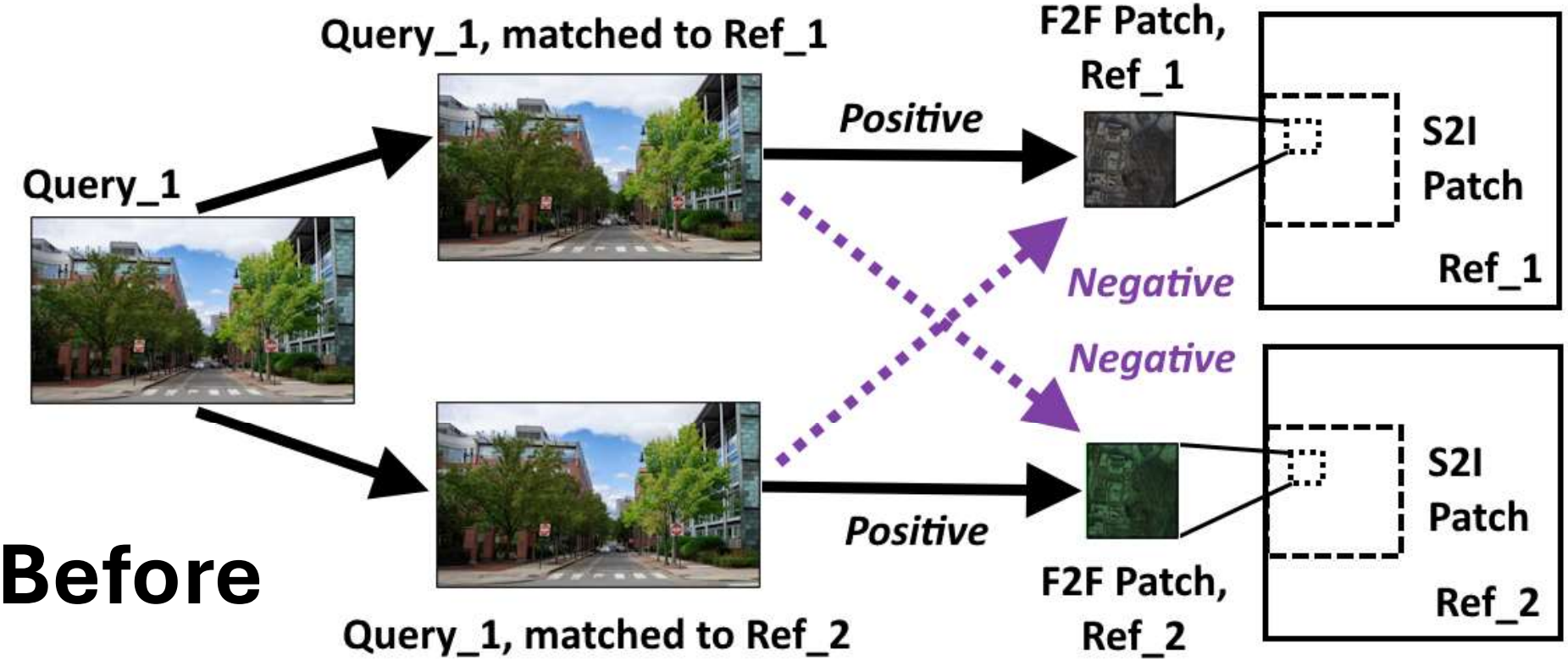
Query videos



Pixel Resolution:
1792 x 1792

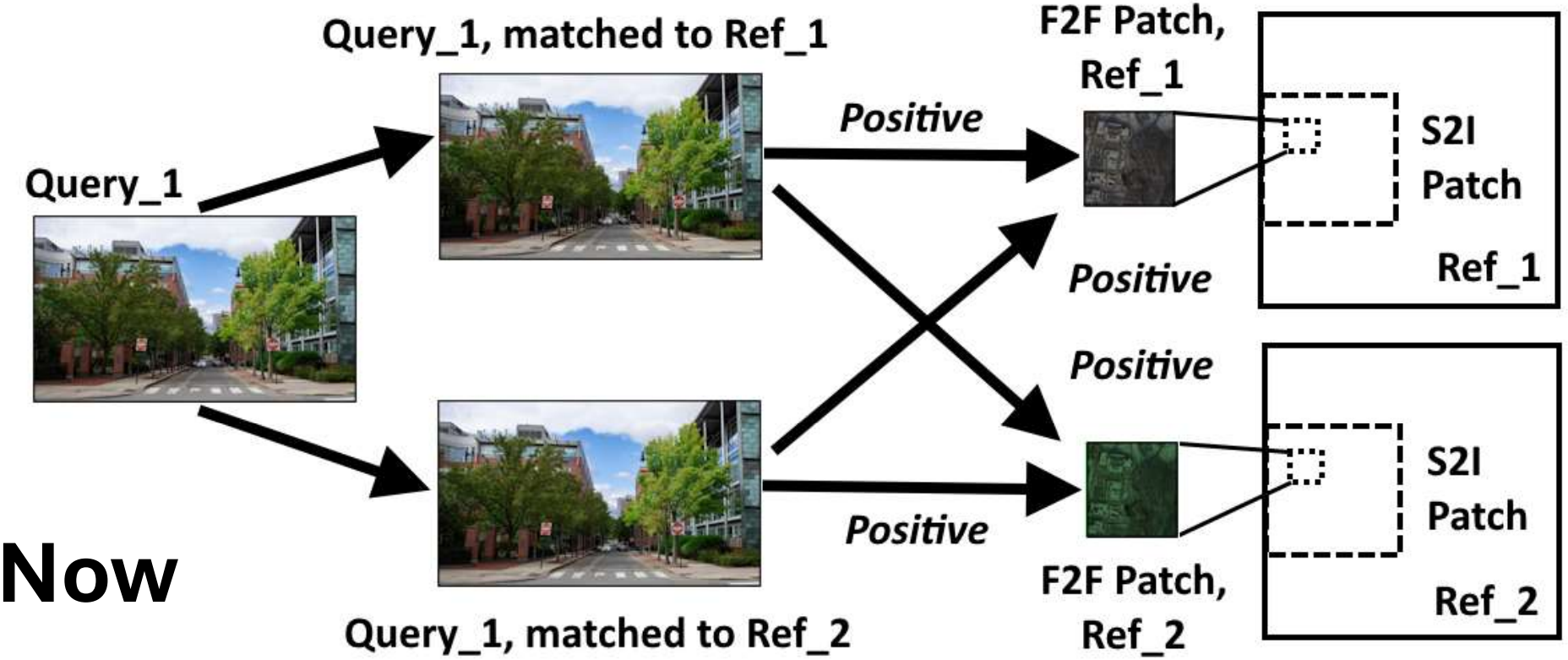
Reference (S2I) Images

GAReT uses single-positive triplet loss for GeoAdapter and the image encoders



Before

Multi-positive triplet loss is better suited for the challenge data



Now

Image encoder fine-tuning procedure

(1) Represent Sites Uniformly

e.g., Sample 3 References per Site:

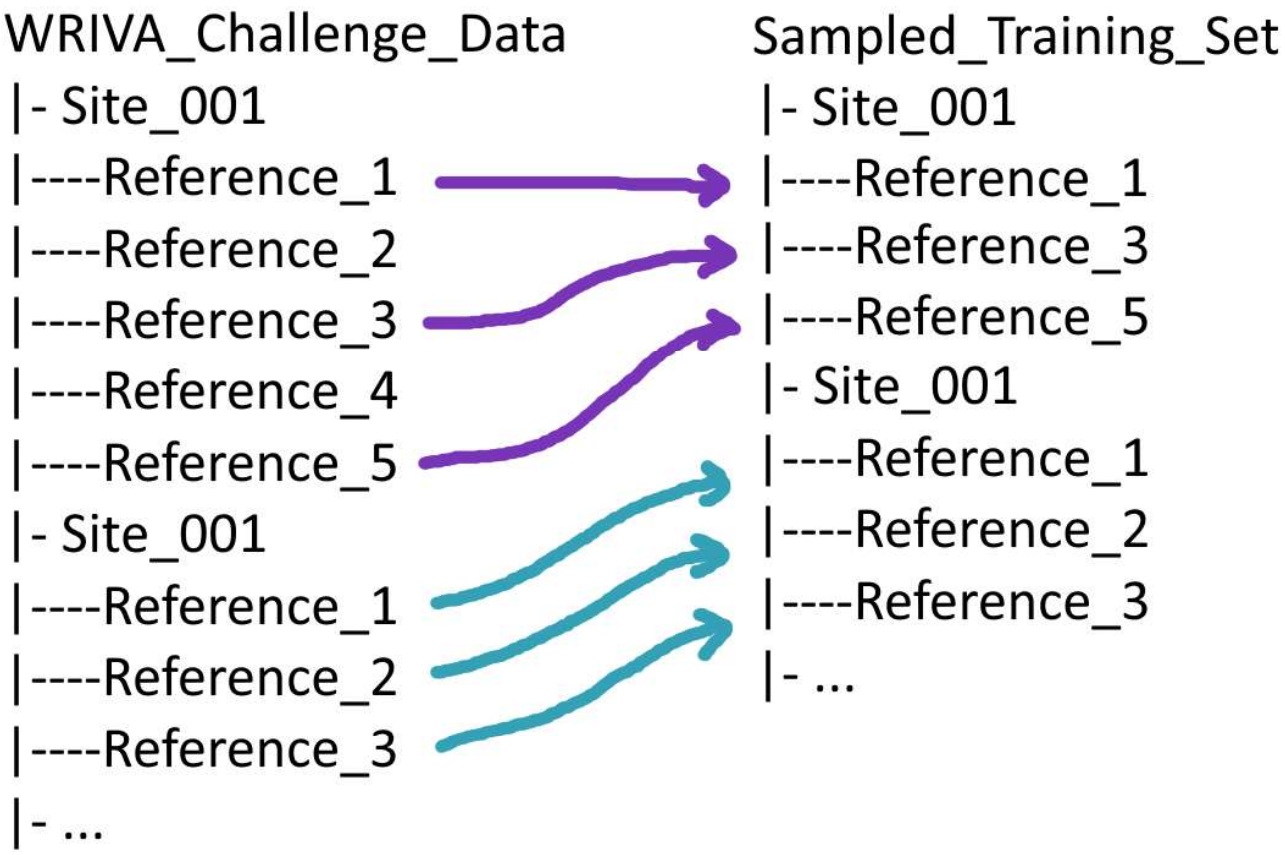
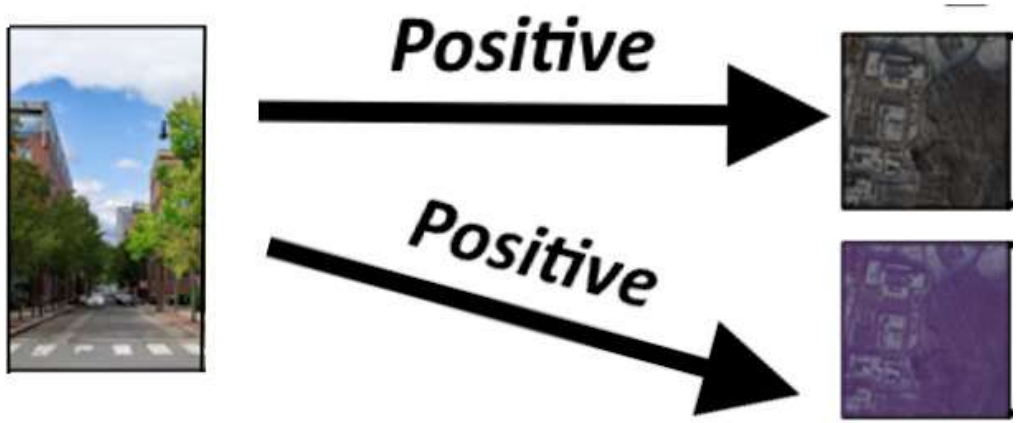


Image encoder fine-tuning procedure

(2) Assign Positives

Corruptions and
Cloud Cover Removed



(3) Load Batches

e.g., Batch Size = 3:

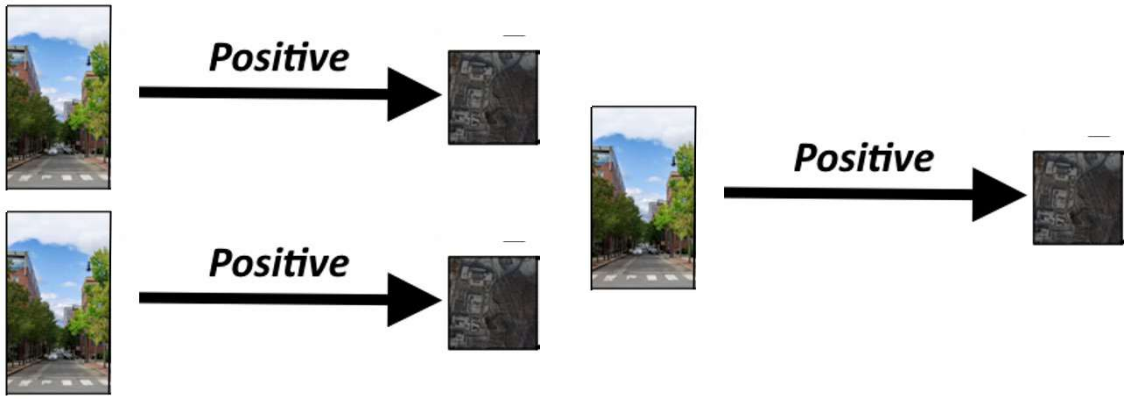
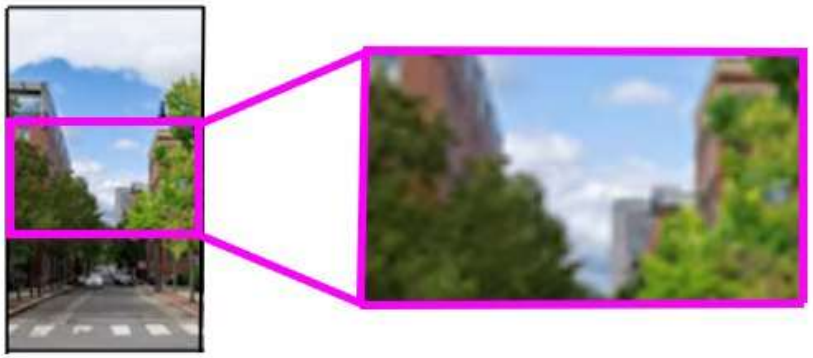


Image encoder fine-tuning procedure

(4) Crop Queries



(5) Encode Batch

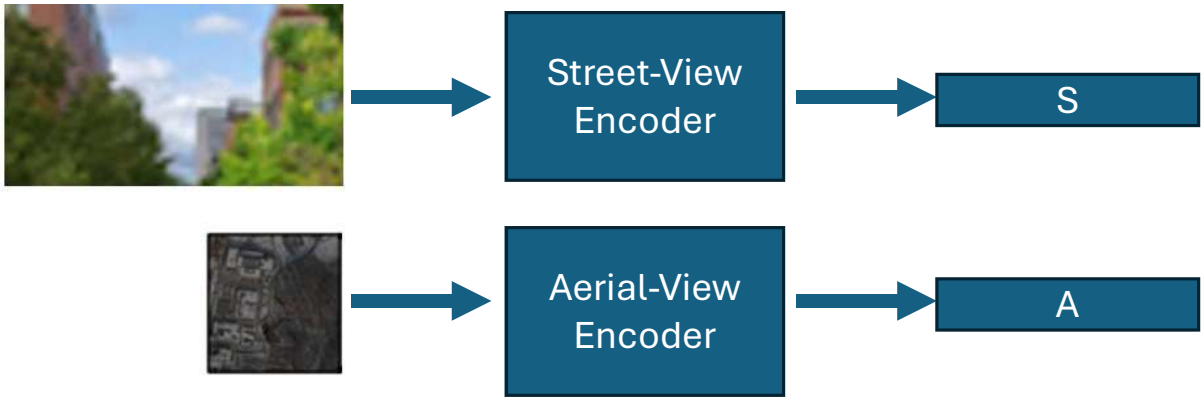
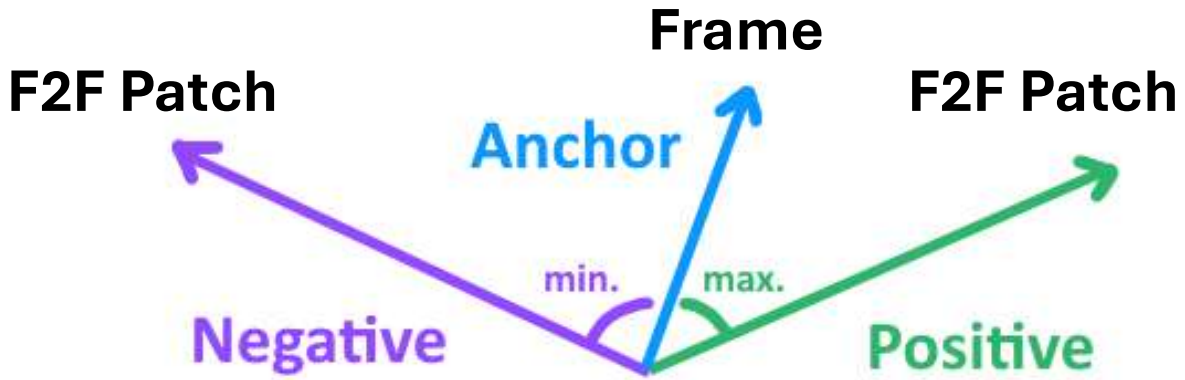


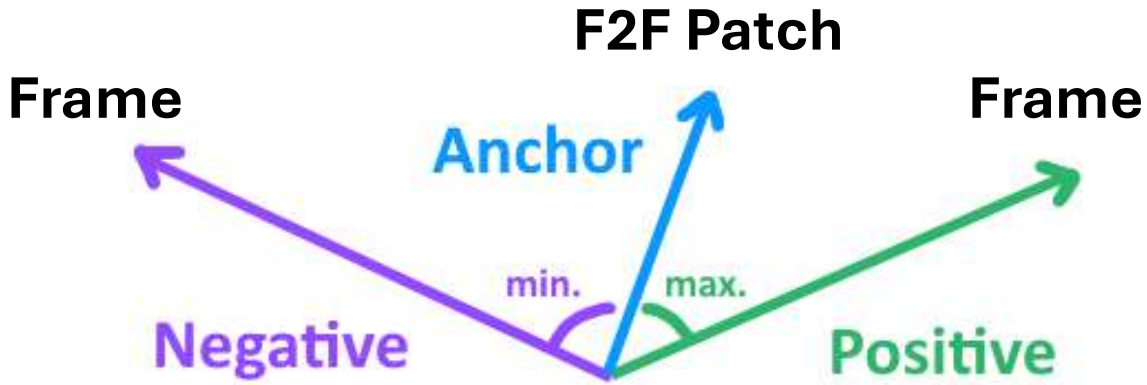
Image encoder fine-tuning procedure

(6) **Bidirectional Soft Triplet Loss**

Direction 1:



Direction 2:



GeoAdapter fine-tuning operates similarly, with the primary difference being positive matching

Query Sequence (8 Frames)



Positive **Negative**

I assume a sequence is always contained by at least one S2I patch



Fine-tuning metrics

106/149 Training Sites, 3/3 Validation Sites

- 6 sampled references / site

512-Pixel S2I Patches (50% Stride), 128-Pixel F2F Patches (50% Stride)

- 49 S2I patches / reference
- 49 F2F patches / S2I

1024-Pixel S2I Patches (50% Stride), 256-Pixel F2F Patches (50% Stride)

- 9 S2I patches / reference
- 49 F2F patches / S2I

Image encoder fine-tuning is overfitting

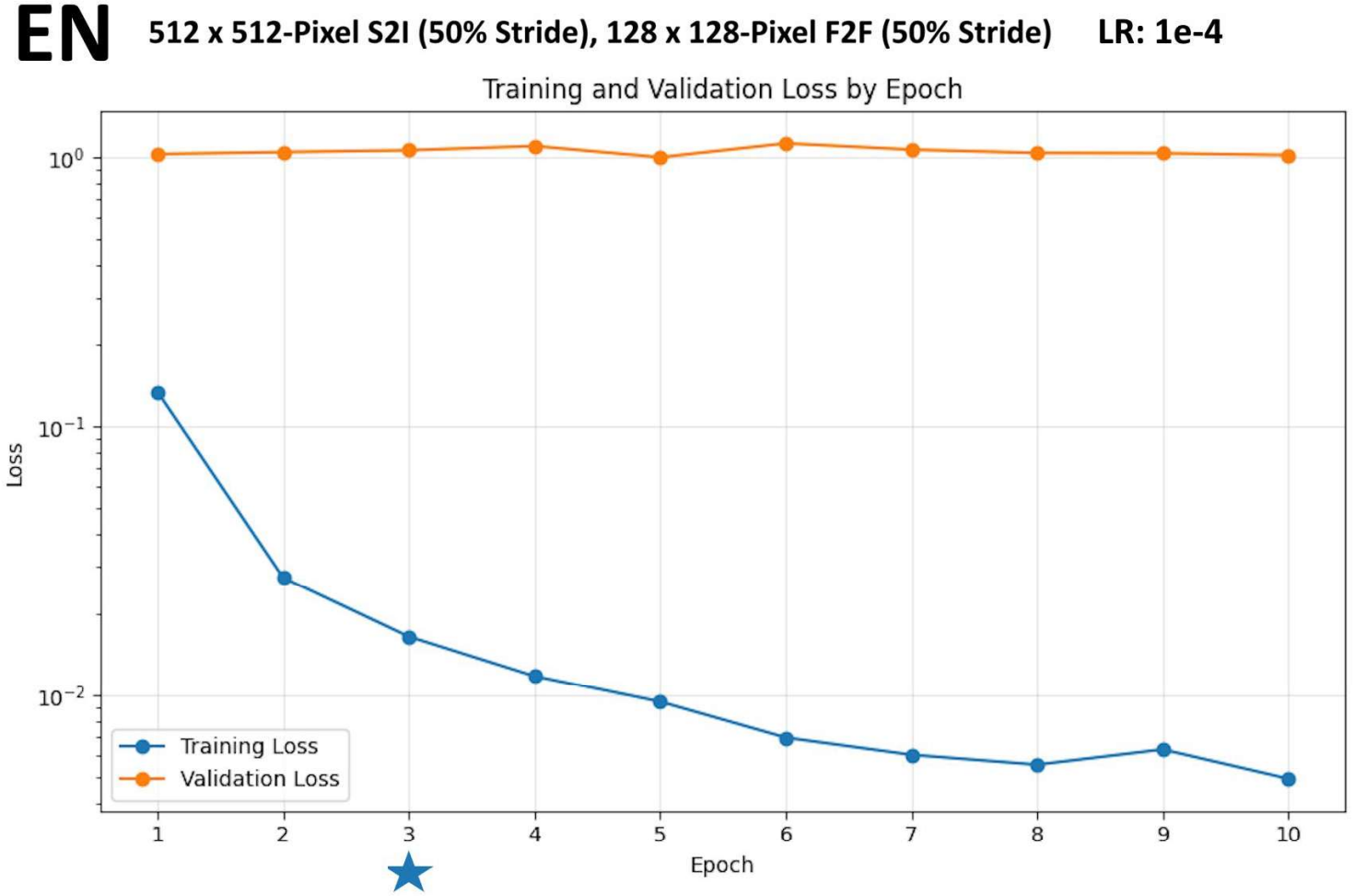


Image encoder fine-tuning is overfitting

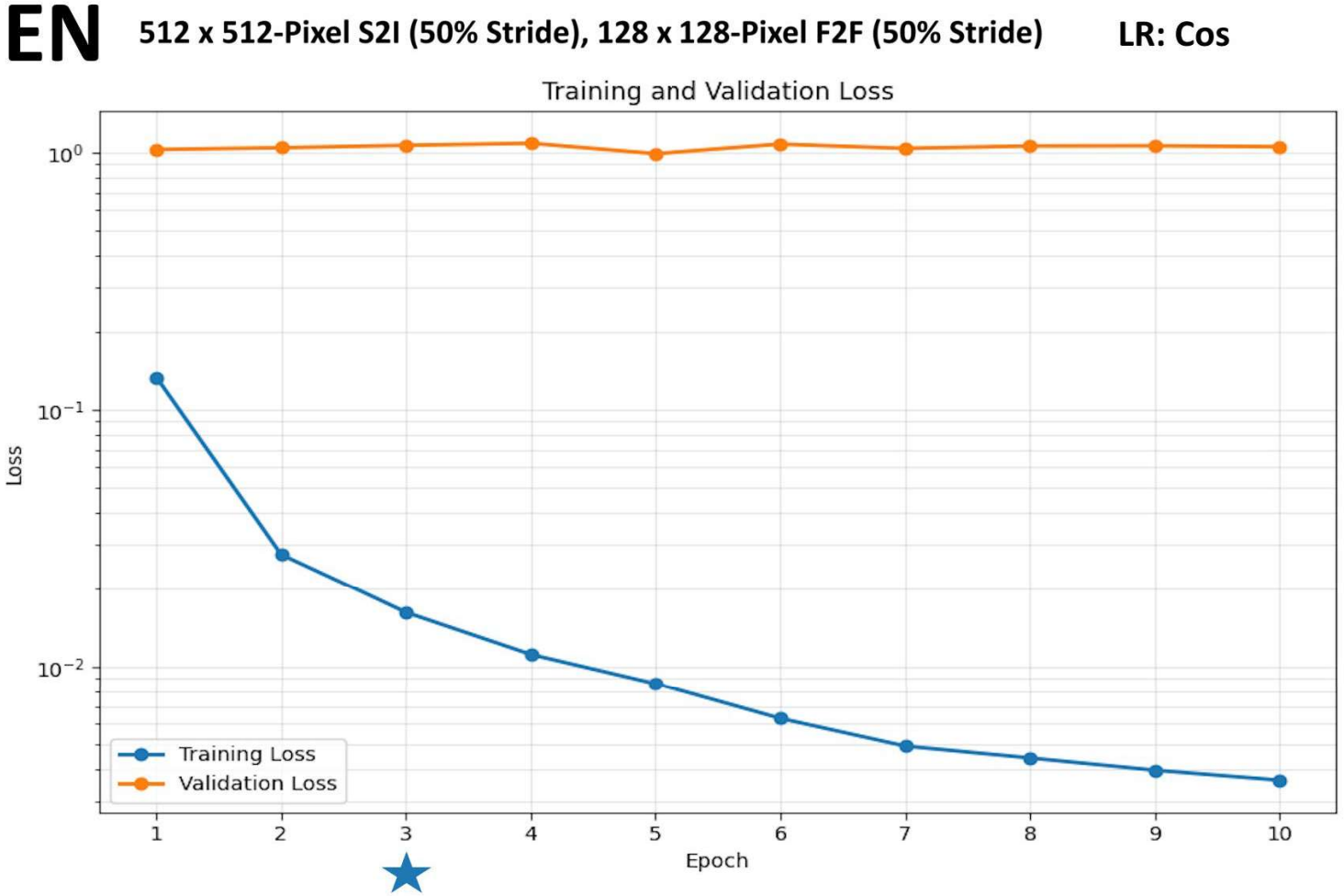


Image encoder fine-tuning is overfitting

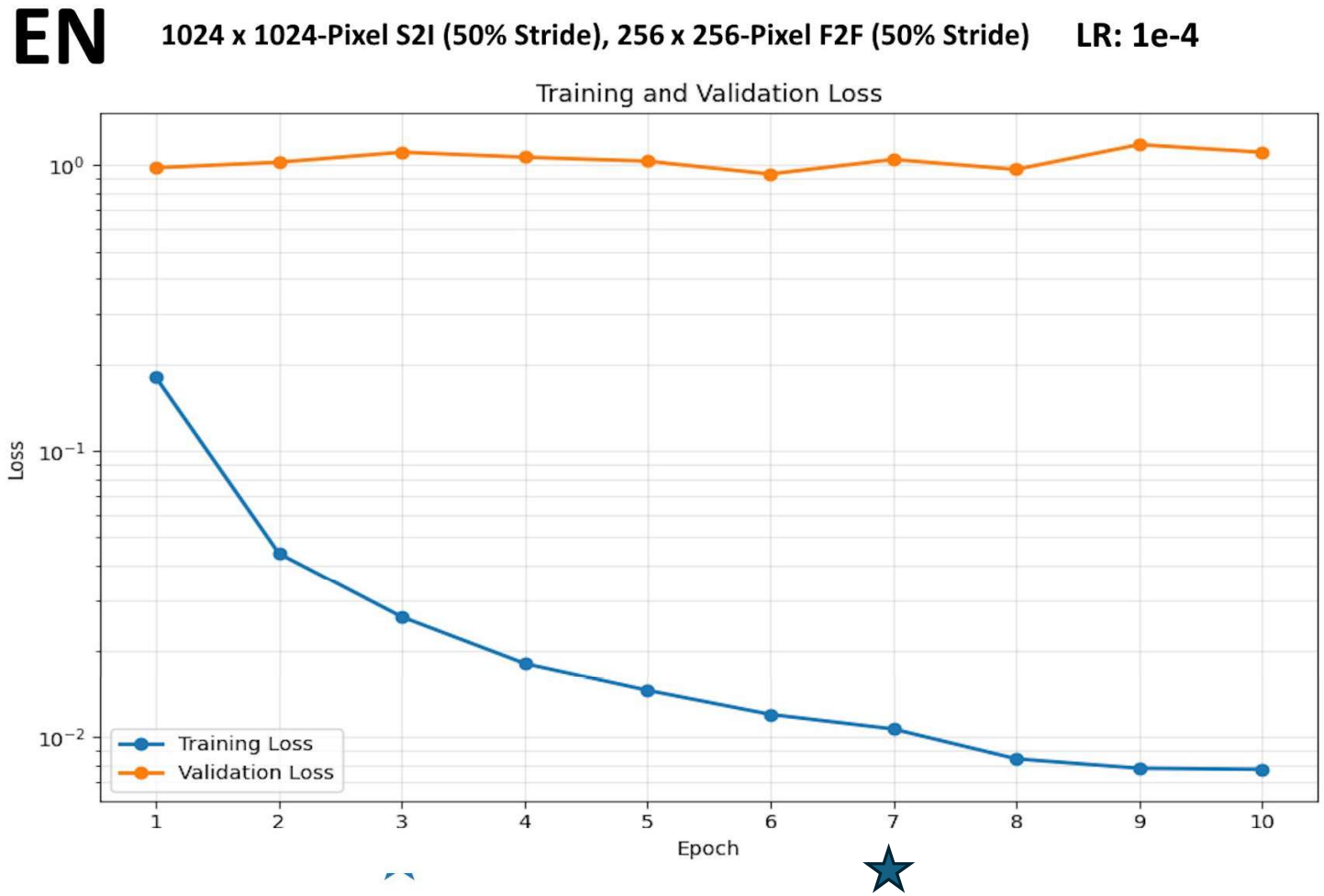
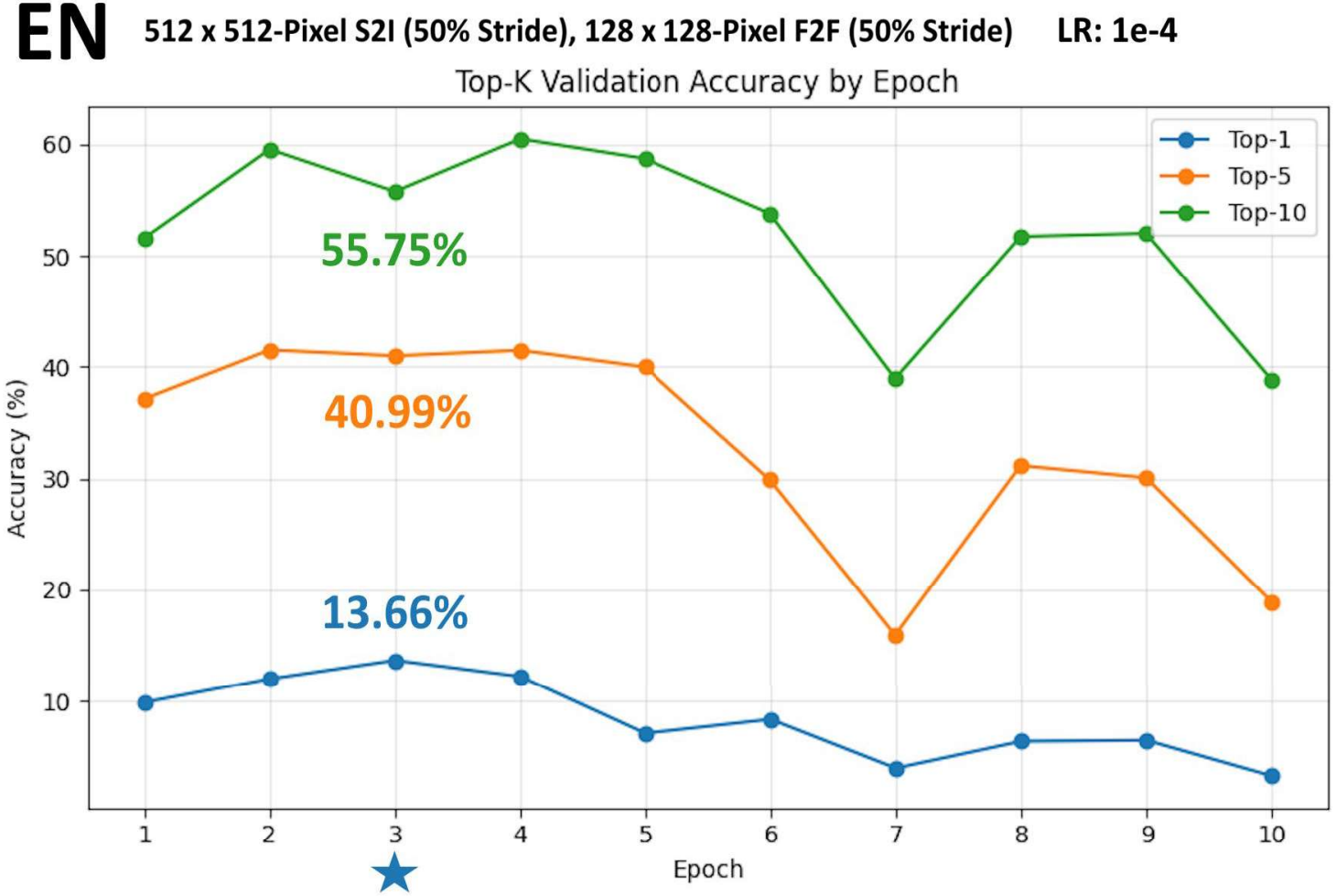


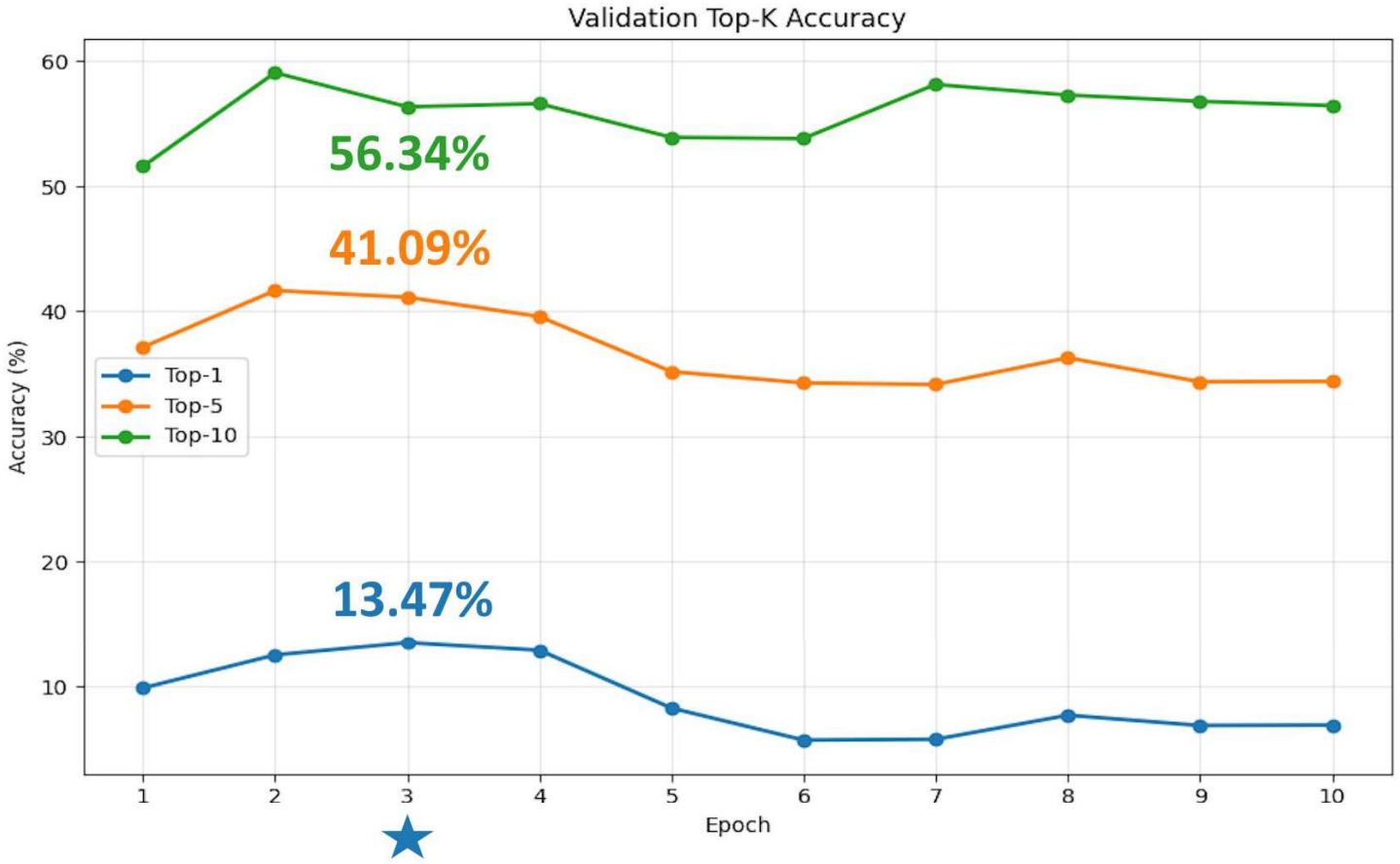
Image encoder fine-tuning is overfitting



Gathers all positive pairings, ranks them using similarity matrix

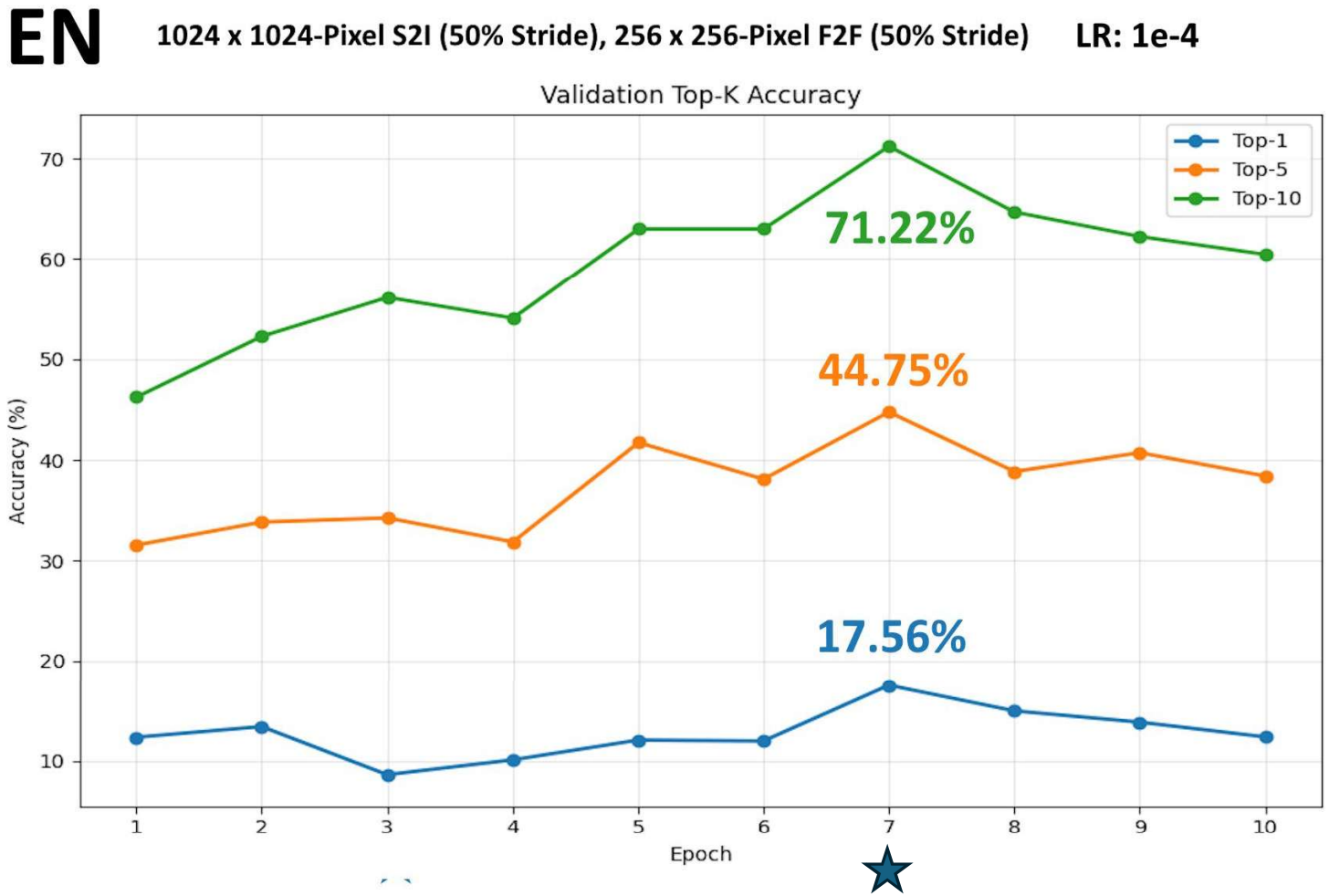
Image encoder fine-tuning is overfitting

EN **512 x 512-Pixel S2I (50% Stride), 128 x 128-Pixel F2F (50% Stride)** **LR: Cos**



Gathers all positive pairings, ranks them using similarity matrix

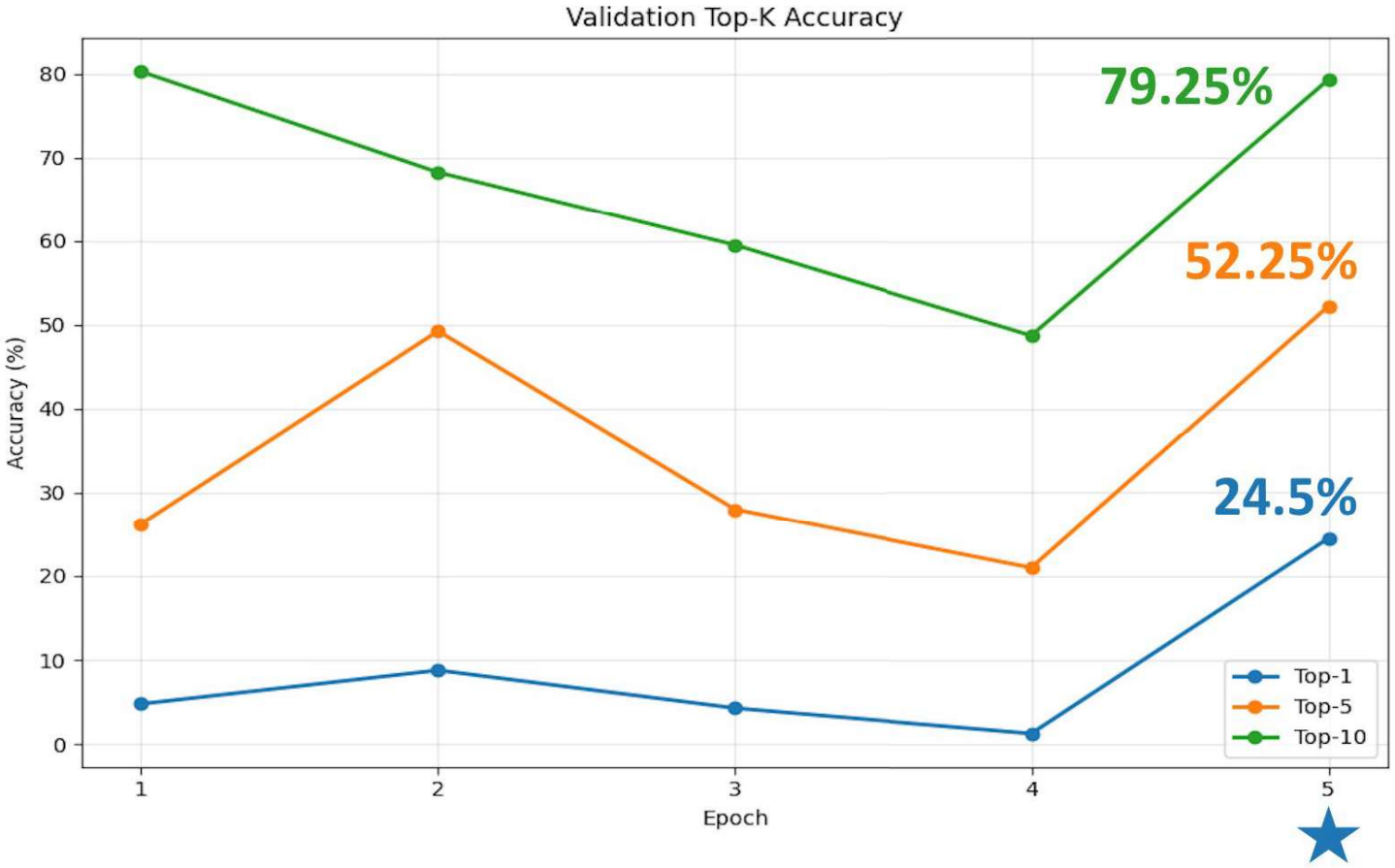
Image encoder fine-tuning is overfitting



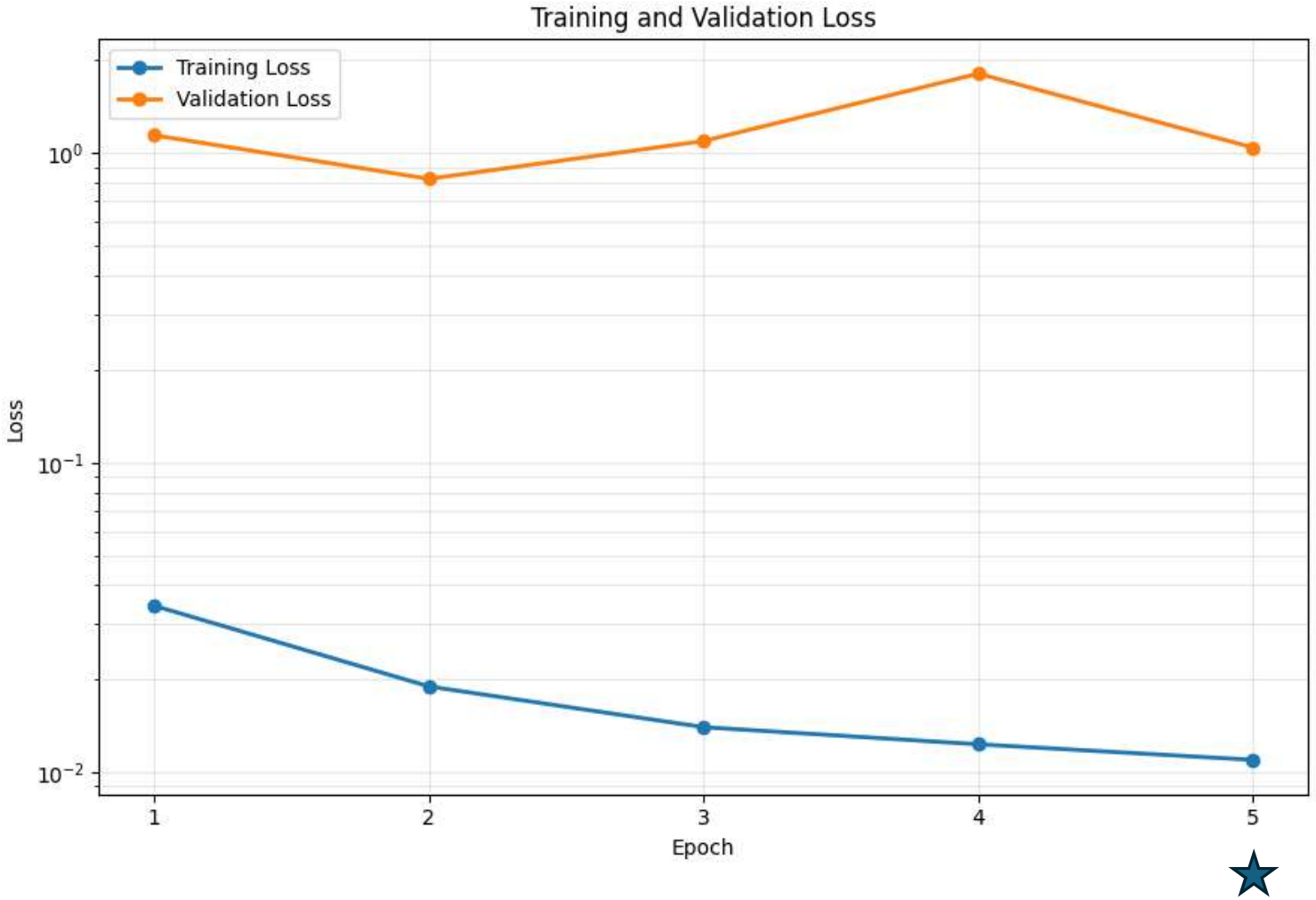
Gathers all positive pairings, ranks them using similarity matrix

GeoAdapter fine-tuning is in-progress

GA **512 x 512-Pixel S2I (50% Stride), 128 x 128-Pixel F2F (50% Stride)** **LR: 1e-4** (Encoders NOT Frozen)

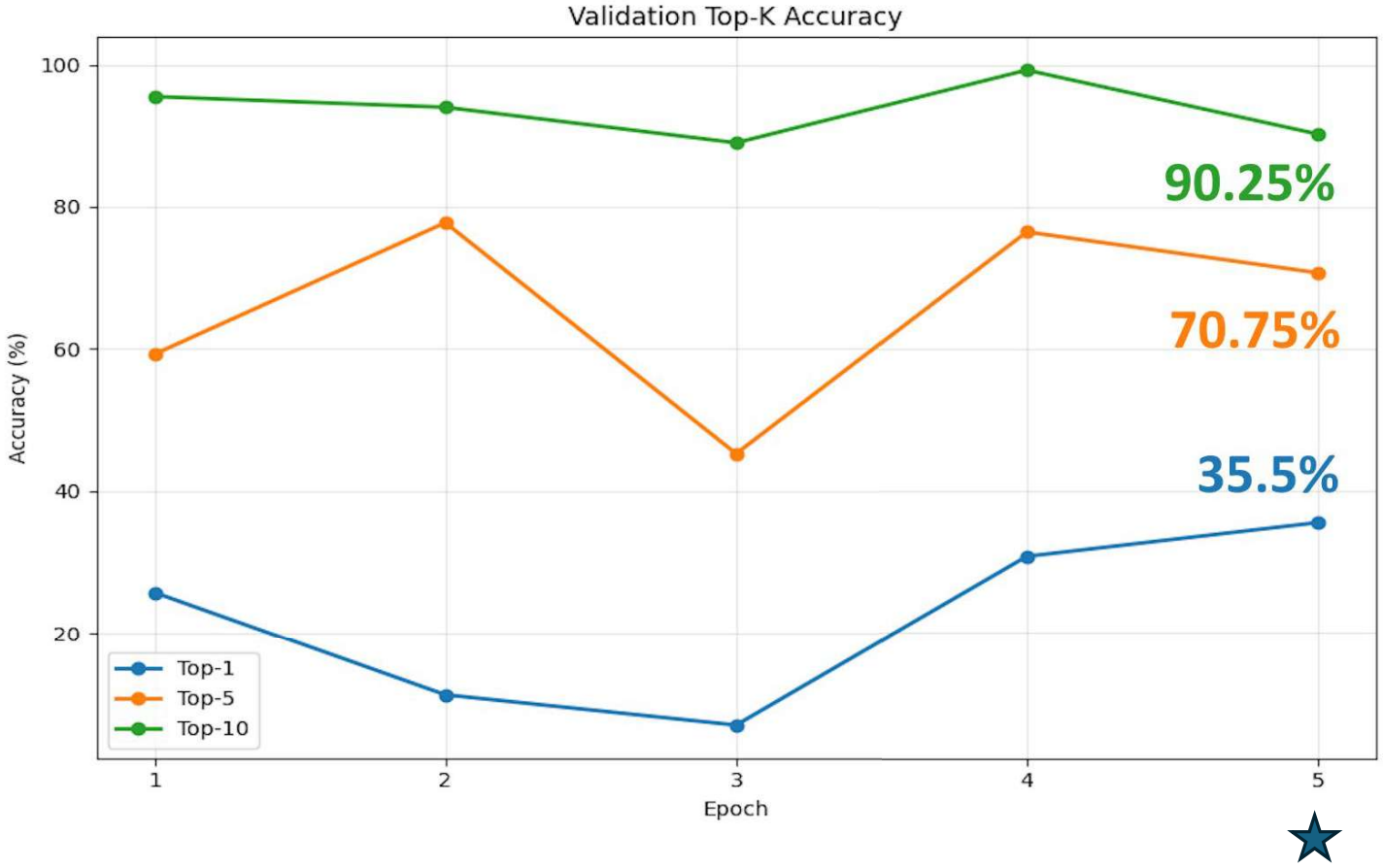


Gathers all positive pairings, ranks them using similarity matrix


Training and Validation Loss
Training Loss
Validation Loss
Loss
10^{0}
10^{-1}
10^{-2}
1
2
3
4
5
Epoch

GeoAdapter fine-tuning is in-progress

GA **1024 x 1024-Pixel S2I (50% Stride), 256 x 256-Pixel F2F (50% Stride)** **LR: 1e-4** (Encoders NOT Frozen)



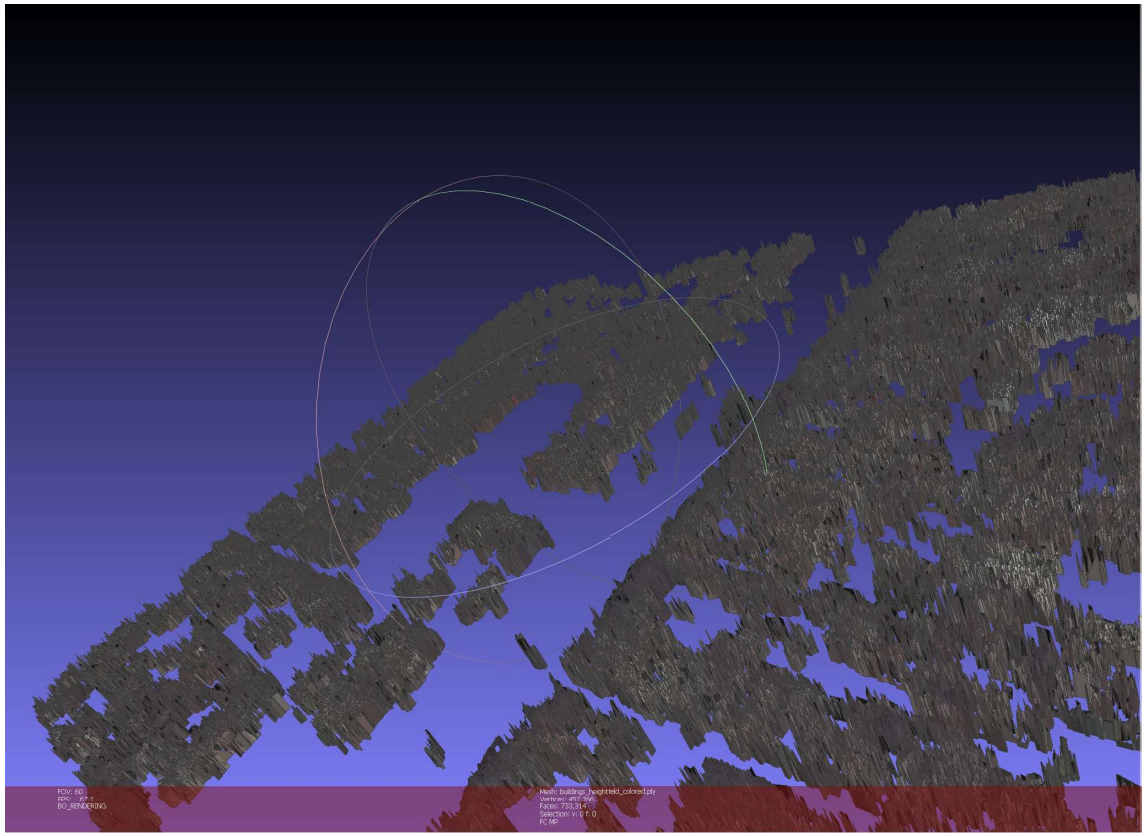
Gathers all positive pairings, ranks them using similarity matrix

Next week's goal: have a fully fine-tuned GAReT adaptation

The reference images invite new 3D approaches



The reference images invite new 3D approaches





Few methods estimate 3D geometry from satellite imagery without RPC metadata