



UCF CENTER FOR RESEARCH
IN COMPUTER VISION

Toward Real-world Cross-view Geo-localization

CVPR 2023 Tutorial: A Comprehensive Tour and Recent
Advancements toward Real-world Visual Geo-Localization

Dr. Chen Chen

Assistant Professor

Center for Research in Computer Vision

University of Central Florida

<https://www.crcv.ucf.edu/chenchen/>

chen.chen@crcv.ucf.edu

Outline

- Introduction (image geo-localization)
- Cross-view image geo-localization
 - **Orientational alignment in image geo-localization**
Sijie Zhu, Taojiannan Yang, Chen Chen, "Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation" Winter Conference on Applications of Computer Vision (WACV), 2021.
 - **Spatial alignment in image geo-localization**
Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
 - **Vision transformer for image geo-localization**
Zhu, Sijie, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Future work

What is image geo-localization?

What's the localization of the place?

Input



Visual Information (Images)

Output



Location in terms of Longitude and Latitude

40.4419, -79.9986

What is image geo-localization?

Query street-view image



GPS location?

(Latitude, Longitude) = (40.441426 , -80.003586)

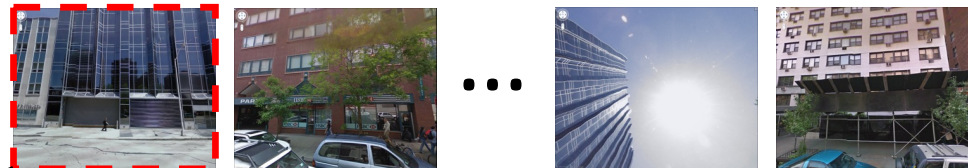


Geo-tagged reference database

Find match



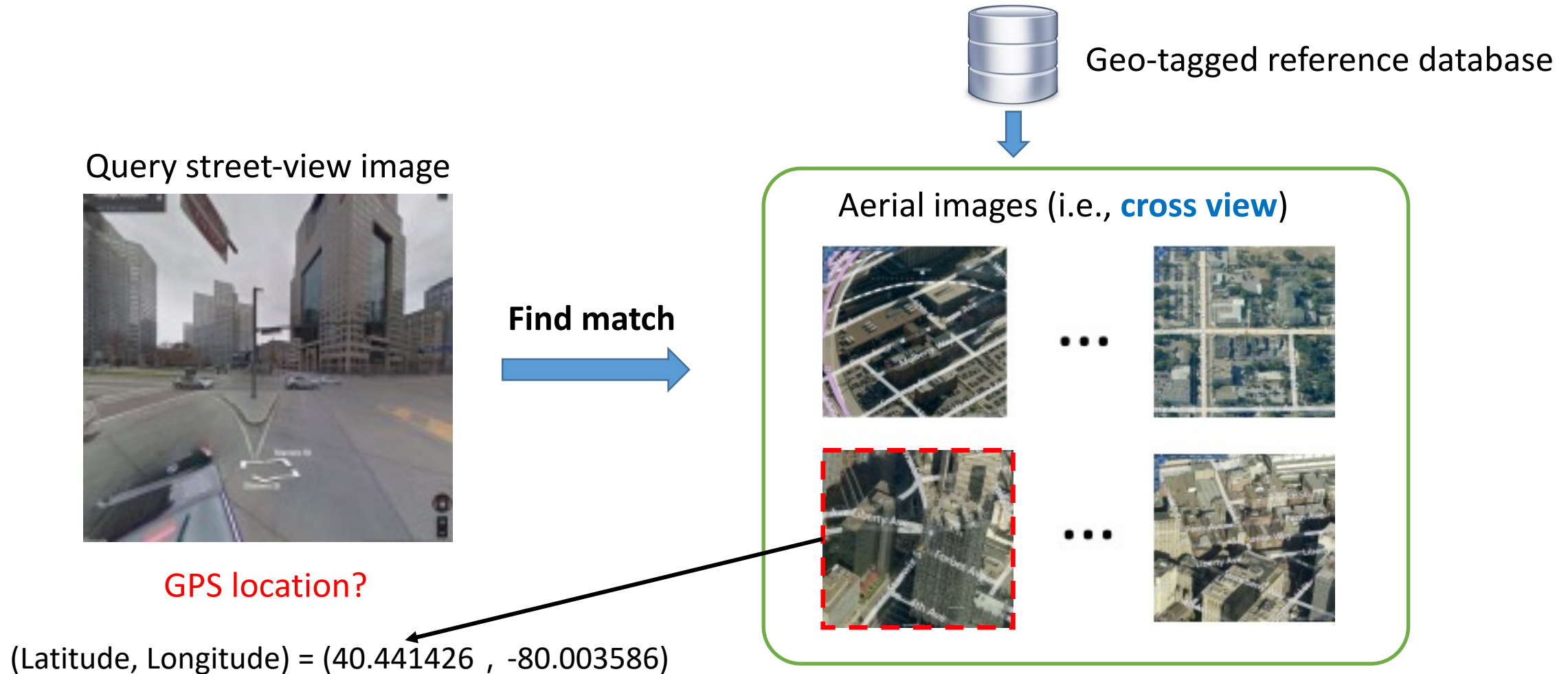
Street-view images (i.e., **same view**)



Top-1 match (ranked by similarity)



What is image geo-localization?



How to define similarity?



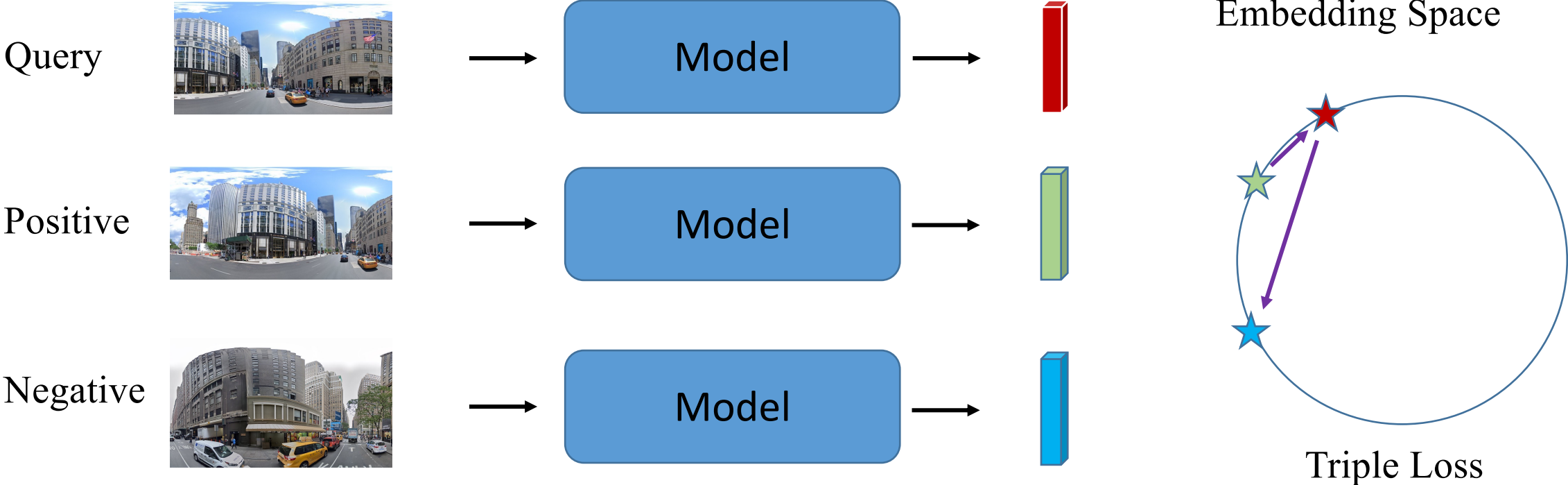
Similar?



Similar?

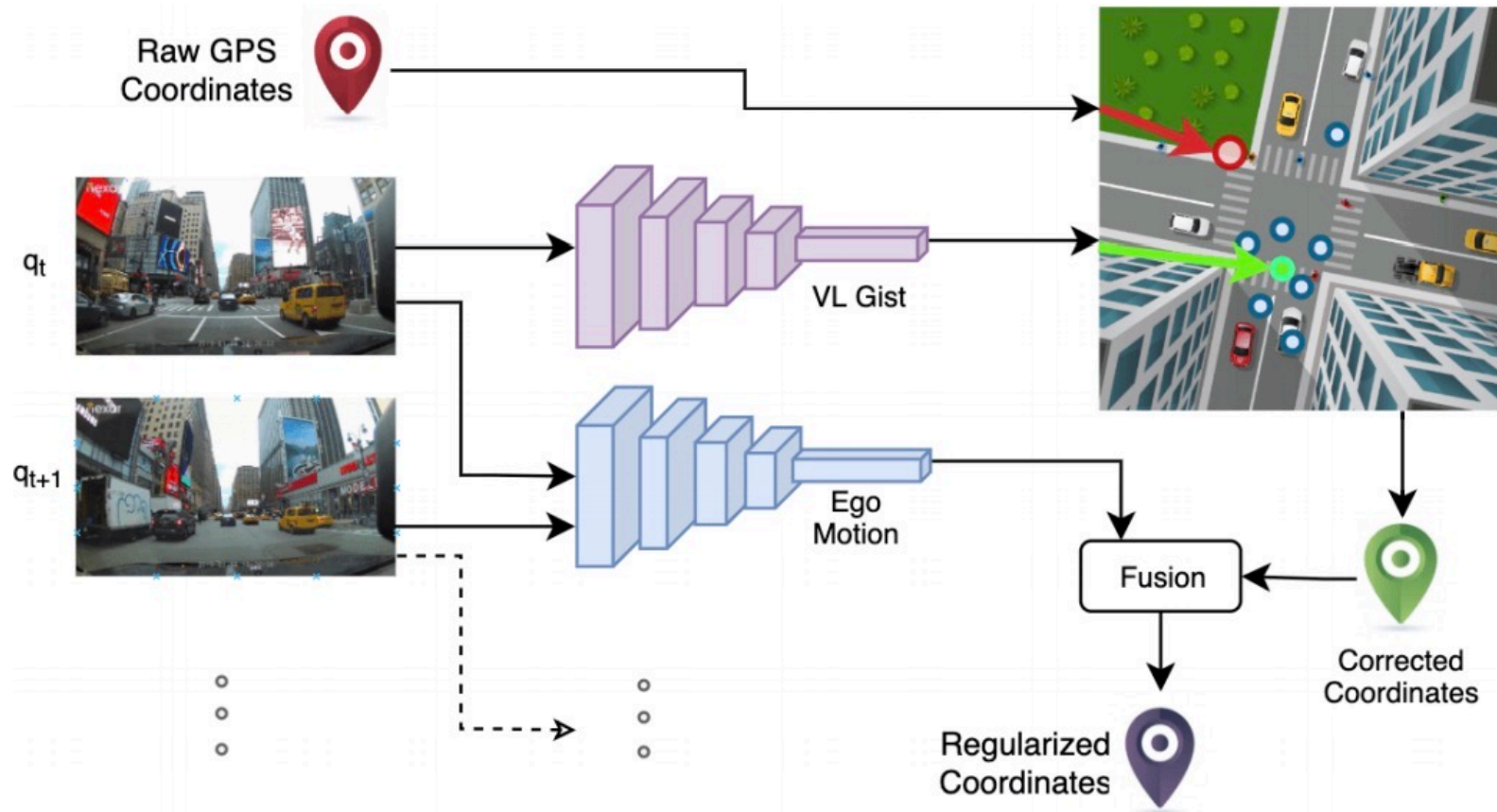


Learn similarity via metric learning



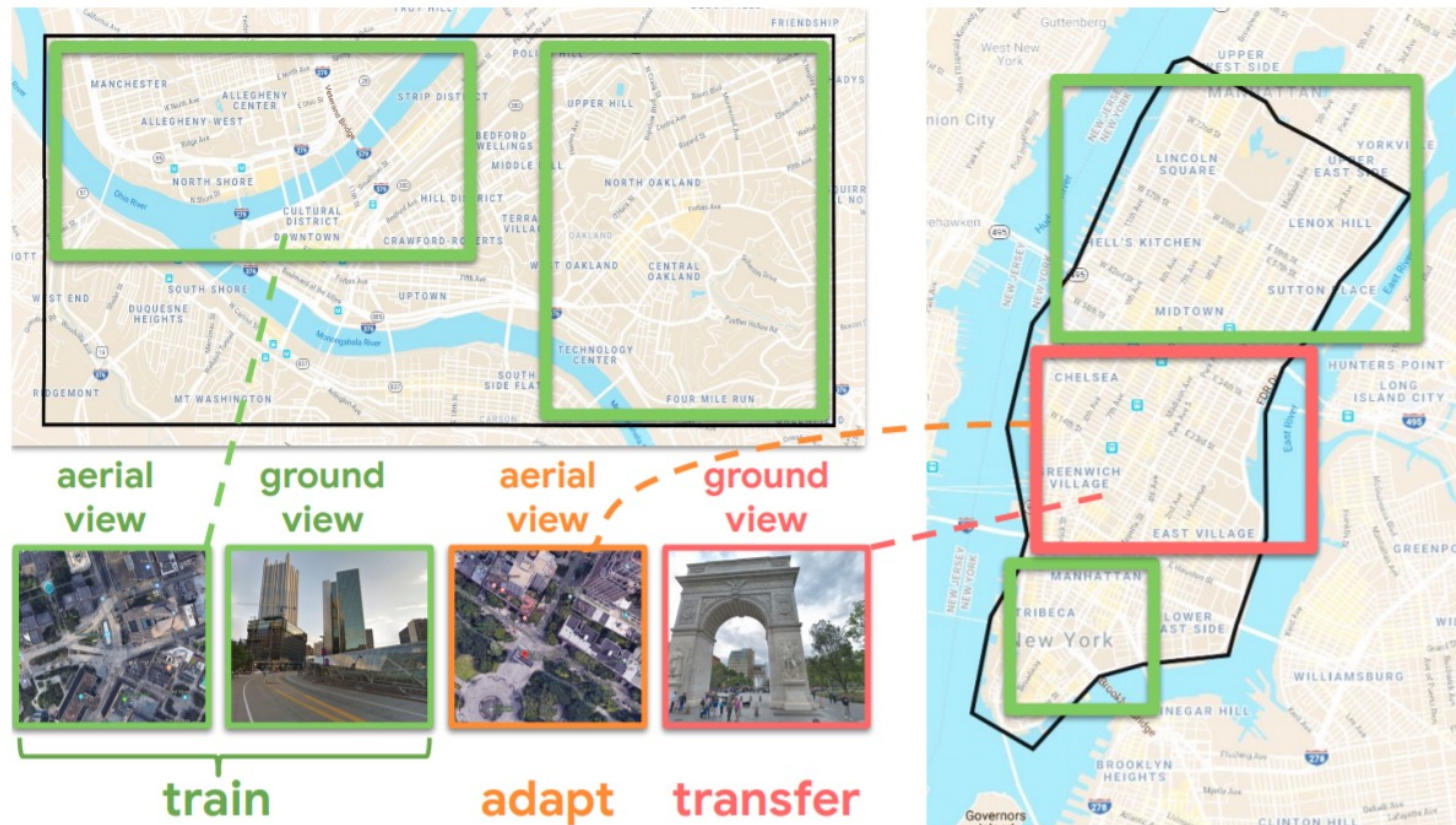
Why is image geo-localization important?

- Accurate Visual Localization for Automotive Applications



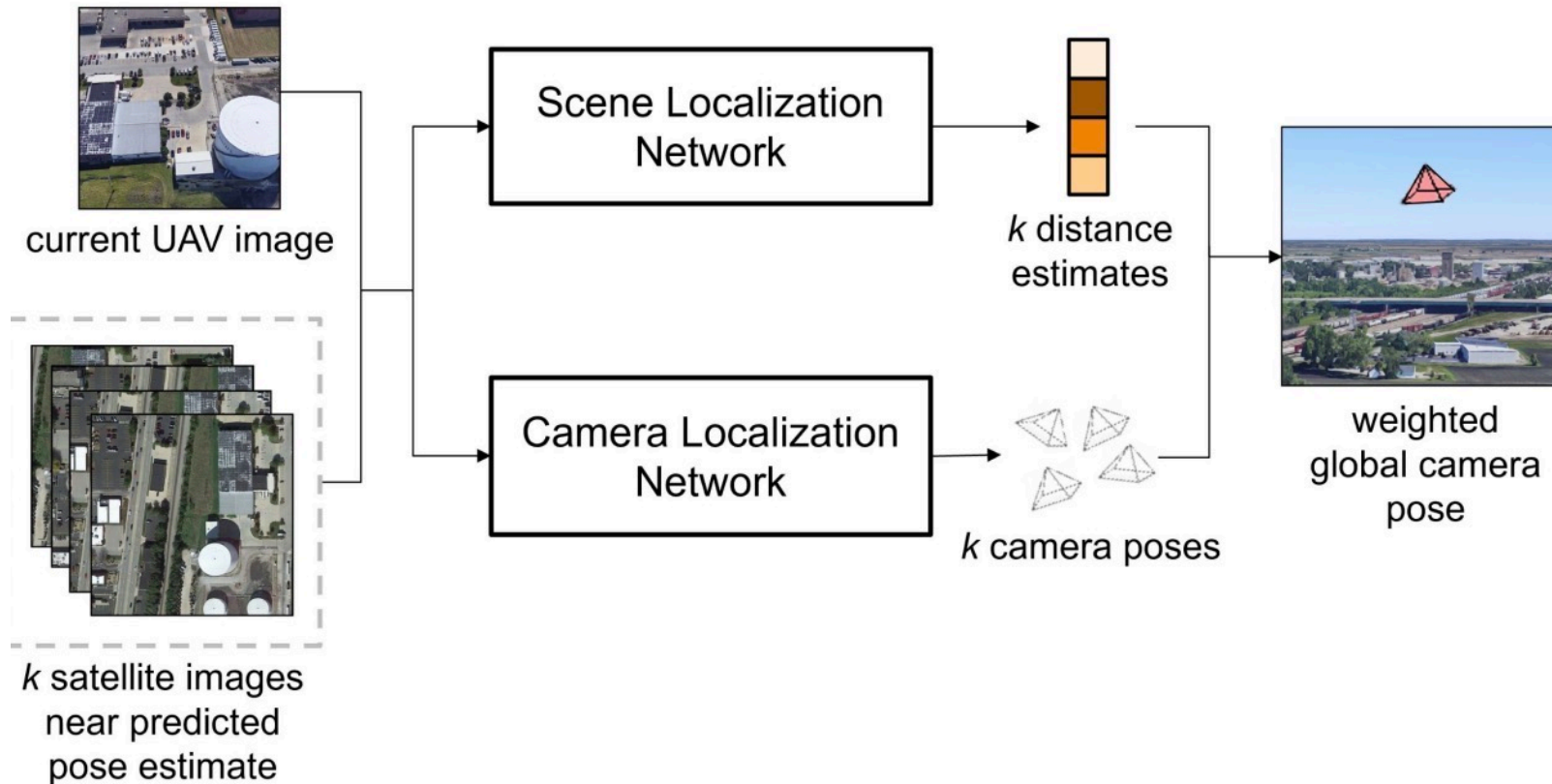
Why is image geo-localization important?

- Cross-View Policy Learning for Street Navigation



Why is image geo-localization important?

- UAV Pose Estimation using Cross-view Geo-localization



Cross-view image geo-localization

- Only small number of cities in the world are covered by ground-level imagery
- A more complete coverage for overhead reference data such as satellite/aerial imagery

Query street-view image



Find match



Geo-tagged reference database

Aerial images (i.e., **cross view**)



...



...



GPS location?

(Latitude, Longitude) = (40.441426 , -80.003586)

Outline

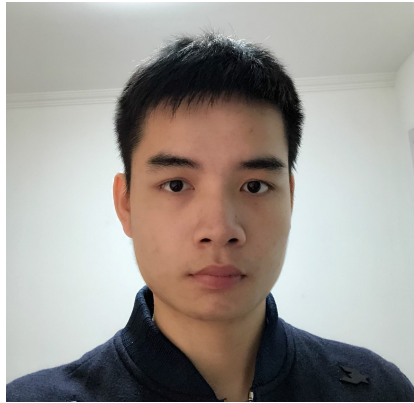
- Introduction (image geo-localization)
- Cross-view image geo-localization
 - Orientational alignment in image geo-localization

Sijie Zhu, Taojiannan Yang, Chen Chen, “Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation” Winter Conference on Applications of Computer Vision (WACV), 2021.
 - Spatial alignment in image geo-localization

Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
 - Vision transformer for image geo-localization

Zhu, Sijie, Mubarak Shah, and Chen Chen. “TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Future work

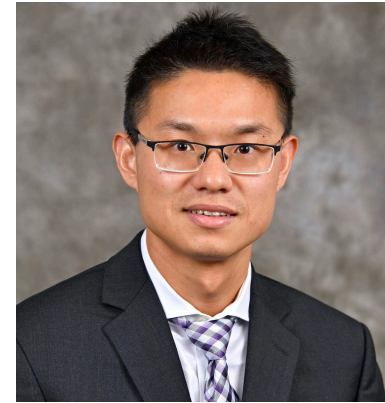
Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation



Sijie Zhu



Taojiannan Yang



Chen Chen

[1] Sijie Zhu, Taojiannan Yang, Chen Chen, “Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation” Winter Conference on Applications of Computer Vision (WACV), 2021.

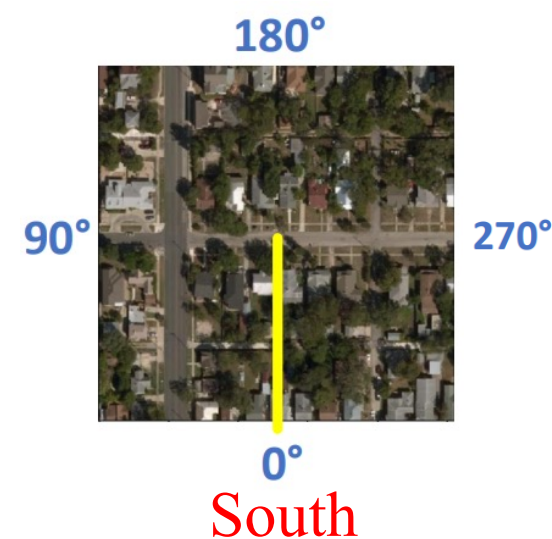
[2] Sijie Zhu, Taojiannan Yang, Chen Chen, “Visual Explanation for Deep Metric Learning”, IEEE Trans. on Image Processing, 2021
<https://arxiv.org/pdf/1909.12977.pdf>

High Accuracy Depends on Assumptions

- Orientation Alignment



South



In real-world applications, images may not be geometrically aligned.

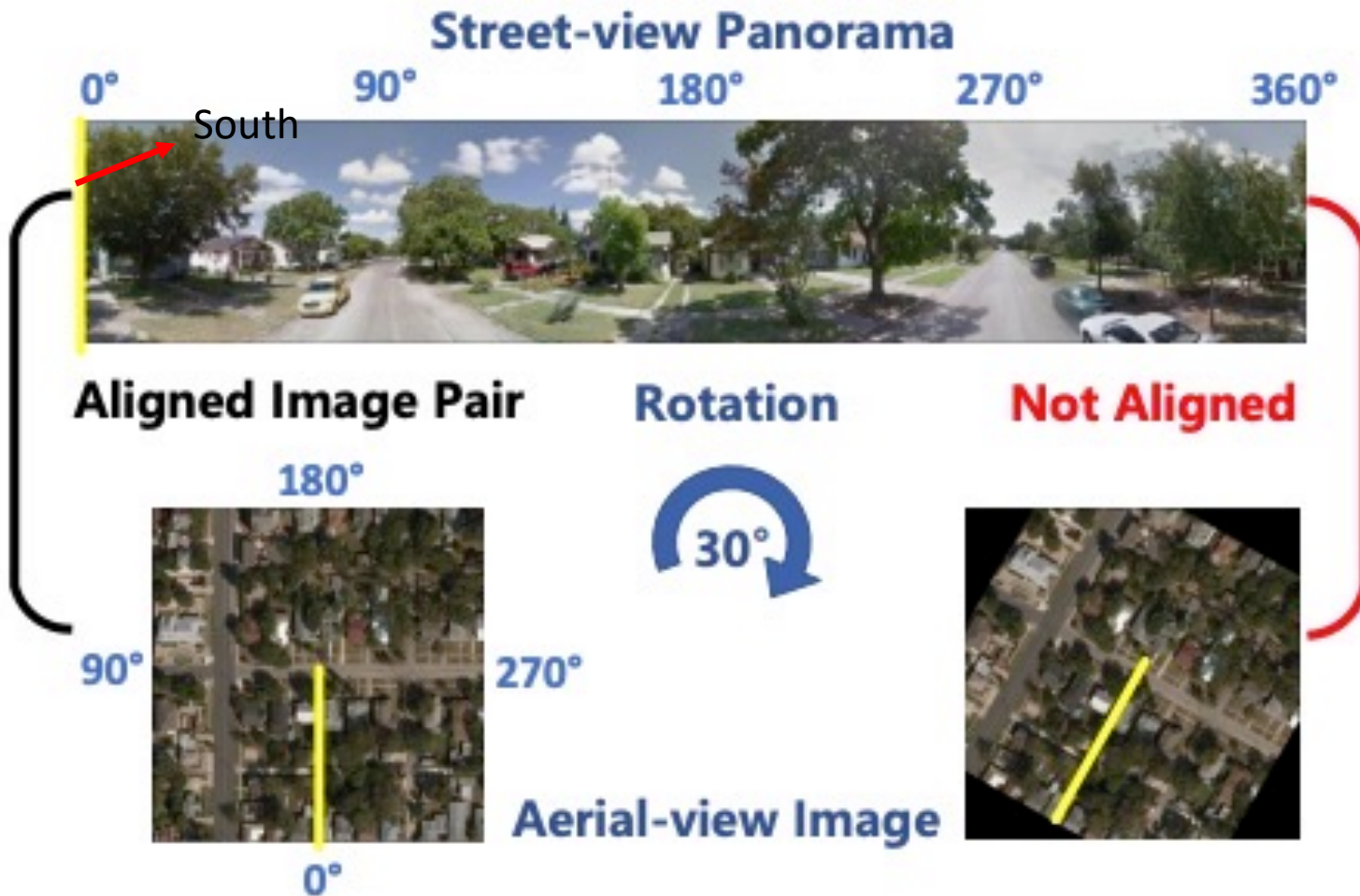
Research problems

- How would the alignment information affect the retrieval model in terms of performance?
- Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?
- Is it possible to estimate the alignment information when no explicit supervision is given?

Research problems

- How would the alignment information affect the retrieval model in terms of performance?
- Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?
- Is it possible to estimate the alignment information when no explicit supervision is given?

Impact of orientation alignment



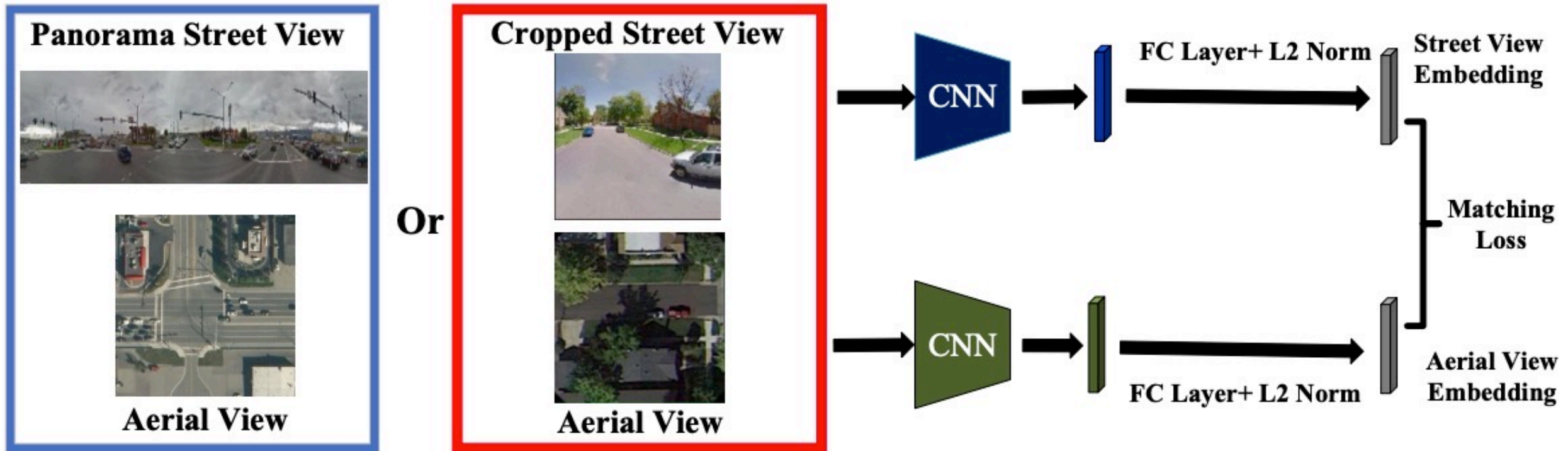
Validation	Training	
	Aligned	Rotate
Aligned	60.1%	43.7%
Rotate	13.5%	44.2%

Top-1 recall accuracy with different alignment settings

Research problems

- How would the alignment information affect the retrieval model in terms of performance?
- Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?
- Is it possible to estimate the alignment information when no explicit supervision is given?

Overall Framework (network architecture)



Metric learning techniques are independent of the alignment assumption

Matching Loss

Triplet loss function

$$L = \frac{1}{N} \sum_i^N \max(0, d_i^p - d_i^n + m)$$

d_i^p and d_i^n denote the distance between the i -th anchor and its positive and negative samples

N : N triplets in a batch

m : a positive margin parameter

Weighted soft-margin loss function

$$L = \frac{1}{N} \sum_i^N \sigma(\alpha(d_i^p - d_i^n)), \quad \alpha > 0$$

soft-margin function $\sigma(d) = \log(1 + \exp(d))$

Matching Loss

Binomial deviance loss (Yi et al.)

$$L = \frac{1}{N_p} \sum_i^{N_p} \sigma(-\alpha(s_i^p - m)) + \frac{1}{N_n} \sum_i^{N_n} \sigma(\alpha(s_i^n - m)).$$

s_i^p and s_i^n denote the cosine similarity between the i -th anchor and its positive and negative samples

N_p and N_n represent the number of positive and negative pairs

m : a positive margin parameter

$$\sigma(d) = \log(1 + \exp(d))$$

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In ICPR, pages 34–39. IEEE, 2014.

Matching Loss

Our new loss function

$$L = \frac{\sum_i^{N_p} \sigma(-\alpha_p (s_i^p - m_p))}{\alpha_p N_p} + \frac{\sum_i^{N_n} \sigma(\alpha_n (s_i^n - m_n))}{\alpha_n N_n}$$

When positive samples are much fewer than negative samples, as in cross-view geo-localization with only one positive match, it would be easier to pulling the only matched sample close to the anchor rather than pushing all negative samples away (i.e., assign a much smaller value to α_p than α_n).

Geo-localization Results

Method	CVUSA		Vo	
	Top-1%	Top-1	Top-1%	Top-1
Scott [22] _(ICCV'15)	34.3%	-	15.4%	-
Zhai [26] _(CVPR'17)	43.2%	-	-	-
Vo [21] _(ECCV'16)	63.7%	-	59.9%	-
CVMNet [8] _(CVPR'18)	93.6%	22.5%	67.9%	-
Lending [13] _(CVPR'19)	93.19%	31.71%	-	-
Reweight [3] _(ICCV'19)	98.3%	46.0%	78.3%	-
GAN [14] _(ICCV'19)	95.98%	48.75%	-	-
Ours	97.7%	54.5%	88.3%	11.8%

Table 2: Top-1 and top-1% recall accuracy comparison on CVUSA and Vo datasets.

“R@k”: If the ground-truth reference image appears in the top k retrieved images, it is considered as correct

Ablation study – effect of alignment

Method	w/ alignment		w/o alignment	
	Top-1%	Top-1	Top-1%	Top-1
Baseline (soft-margin loss)	98.8%	60.1%	96.9%	43.7%
Ours (proposed loss)	99.1%	70.4%	97.7%	54.5%

w/o alignment: images are randomly rotated in the training and test sets

- The improvements of the proposed technique are consistent across both settings

Geo-localization Examples



Street view id:6312,file name:0023612.jpg

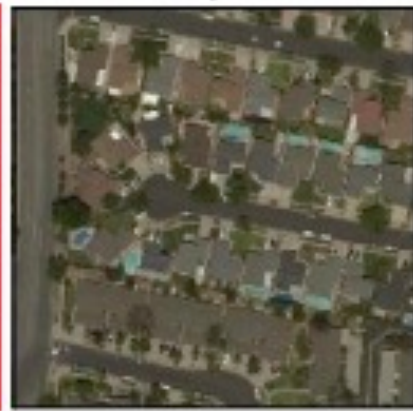
Aerial view rank:1



Top-1



Top-2



Top-3



Top-4



Top-5



Geo-localization Examples

Street view id:18110



Aerial view rank:1
similarity:0.76



Top-1
similarity:0.76



Top-2
similarity:0.75



Top-3
similarity:0.74



Geo-localization Examples

A failure case

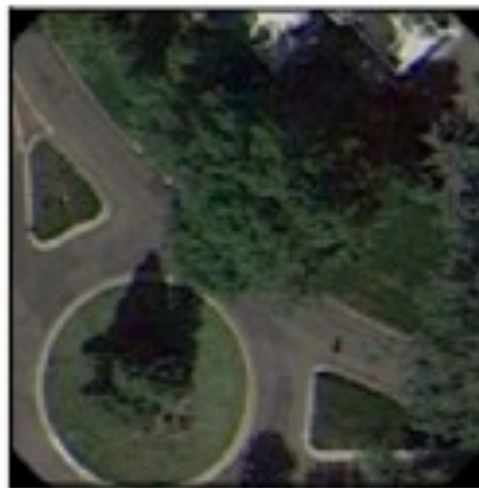
Street view id:63013



Aerial view rank:5984
similarity:0.47



Top-1
similarity:0.78



Top-2
similarity:0.76



Top-3
similarity:0.76



Visual Explanation of the Matching Results

- Visual explanation using Grad-CAM

**Ground Truth
Aerial Image**

Query



Positive Pair

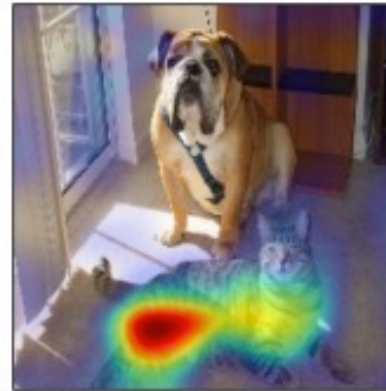


What is Grad-CAM?

- Gradient-weighted Class Activation Mapping (Grad-CAM)



(a) Original Image



(c) Grad-CAM 'Cat'



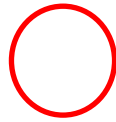
(i) Grad-CAM 'Dog'

Visual Explanation of the Matching Results

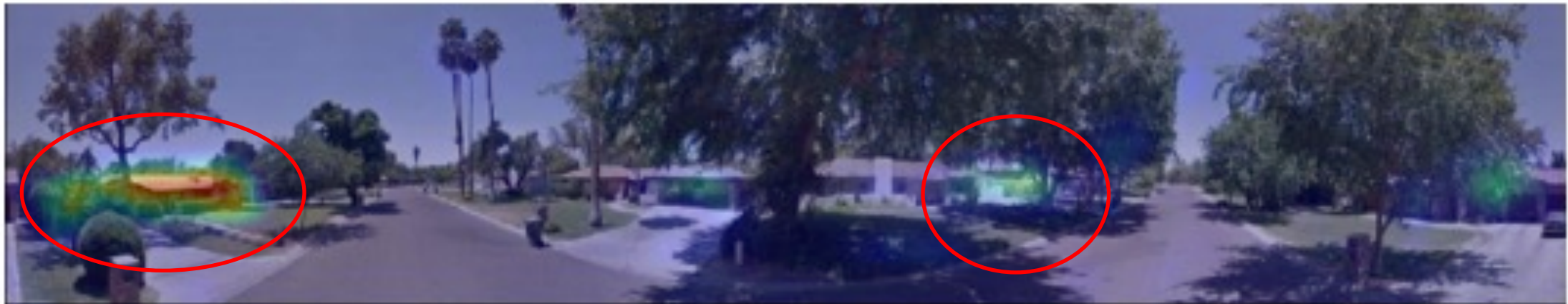
Ground Truth
Aerial Image



Positive Pair

 : regions that contribute the most to the similarity measure

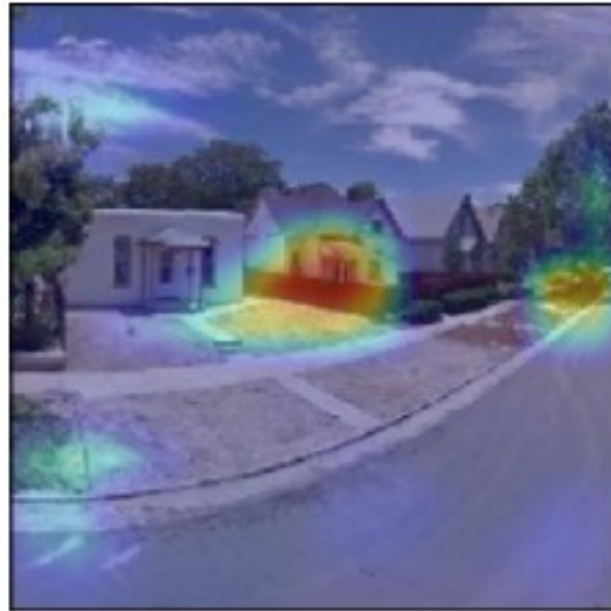
Query



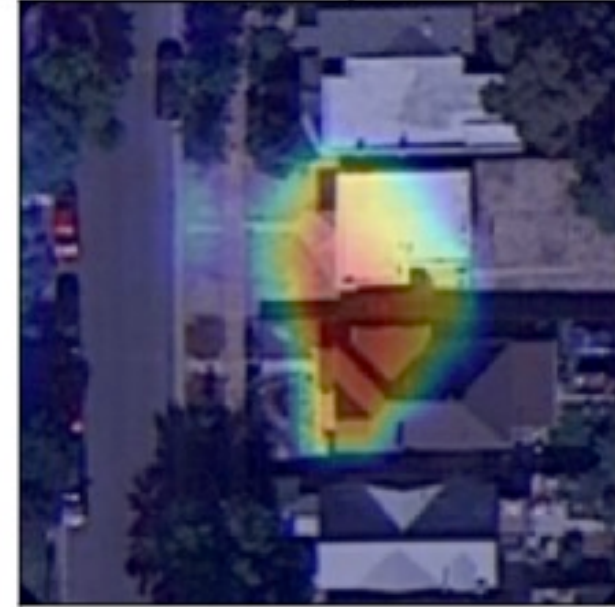
The most activated regions are likely to be the same objects

Visual Explanation of the Matching Results

Street view id:49288



Aerial view positive, id:49288

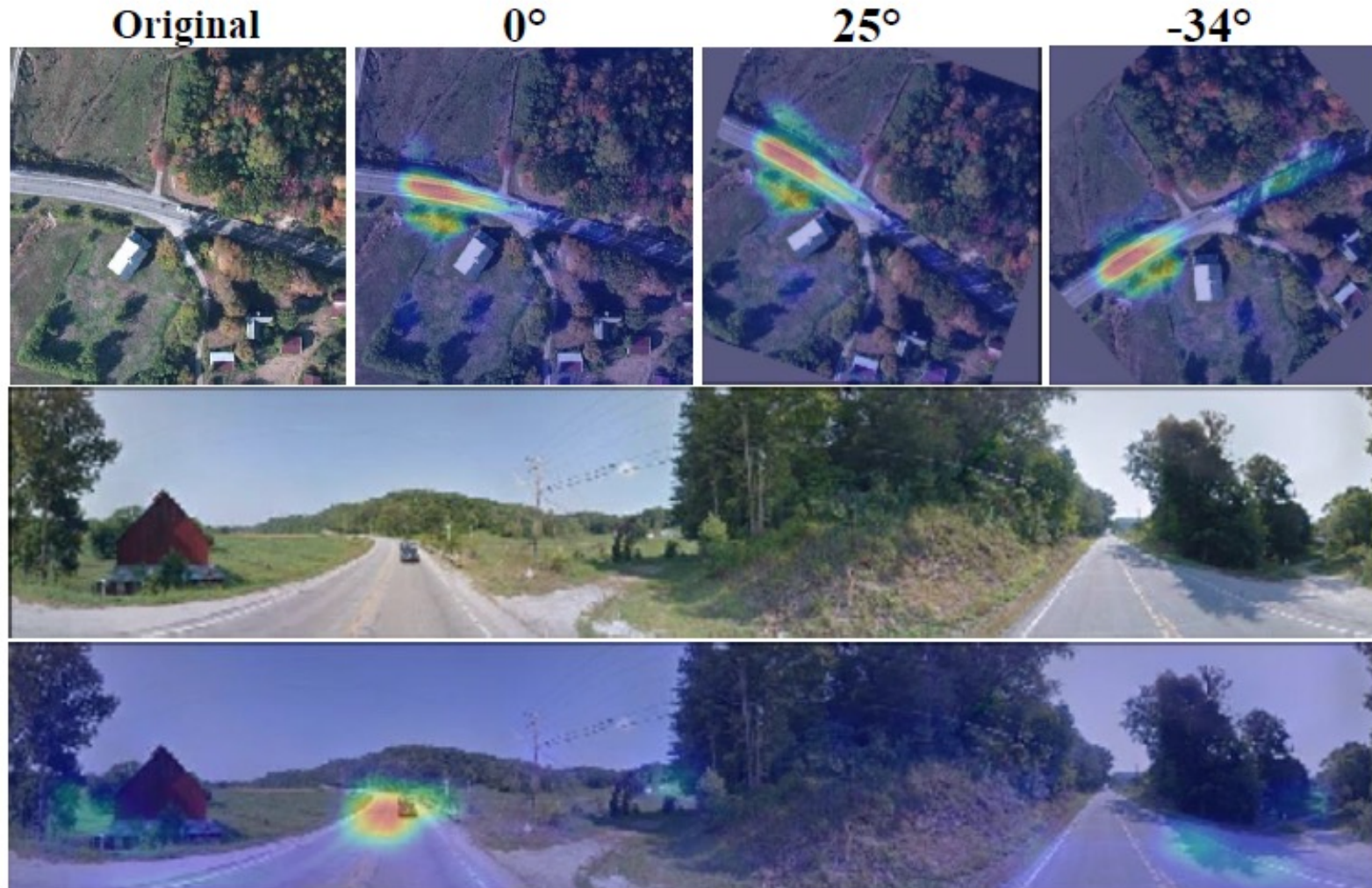


Similarity:0.74

Research problems

- How would the alignment information affect the retrieval model in terms of performance?
- Without assuming the inference image pairs are aligned, how to effectively improve the retrieval performance?
- Is it possible to estimate the alignment information when no explicit supervision is given?

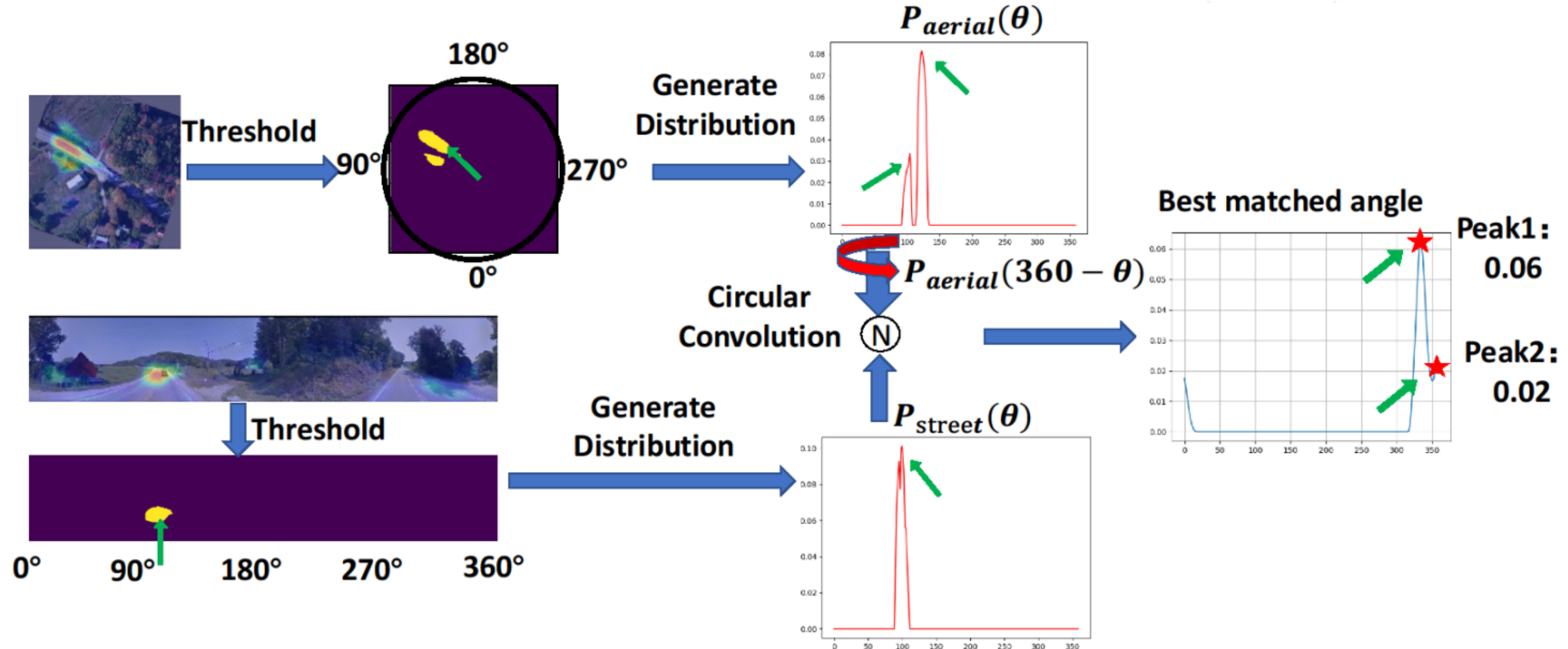
Orientation Estimation with Grad-CAM



We find the Grad-CAM activation maps have the rotation-invariant property!

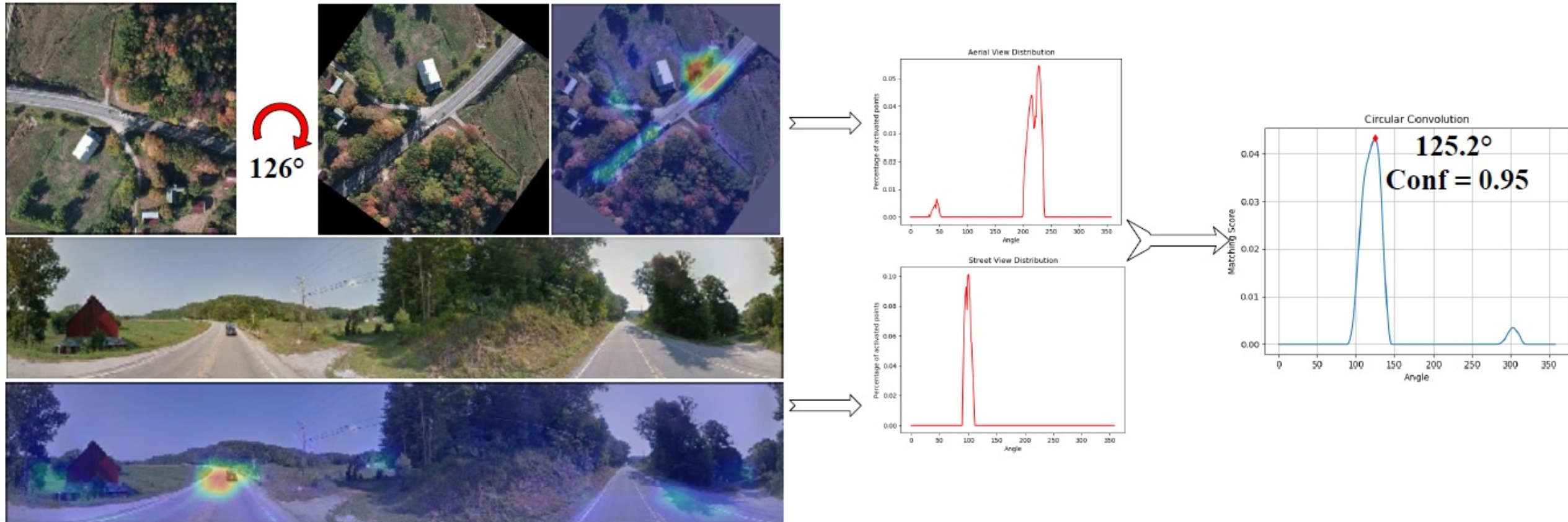
Rotation Estimation Pipeline

find the angle ϕ so that $p_{aerial}(\theta + \phi)$ best matches $p_{street}(\theta)$



The angle distributions of activated pixels from two views would be similar if the image pair is well aligned.

Orientation Estimation Example

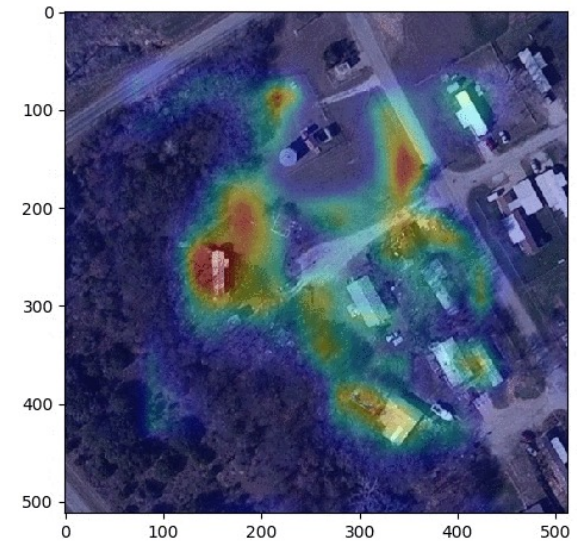
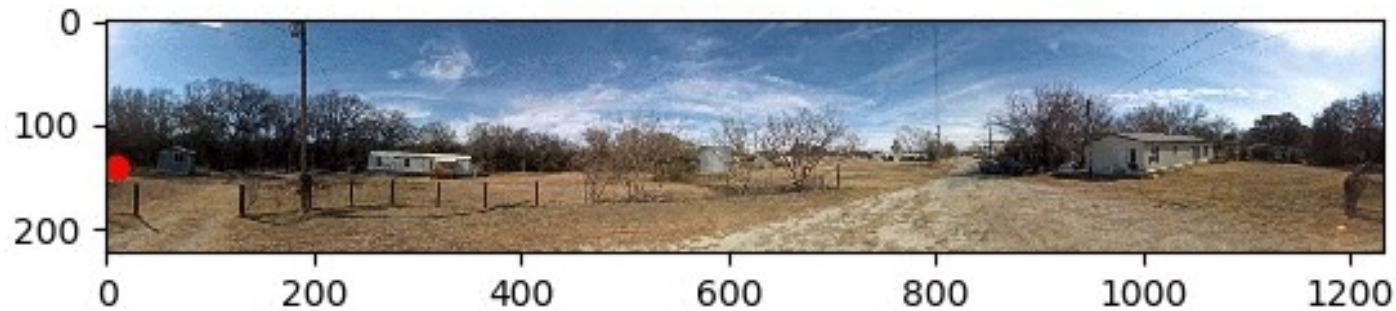


Summary

- Ablation study and visual explanation lead to a key observation – the orientation alignment has a great impact on the retrieval performance (overlooked by prior work)
- We show that improvements on metric learning techniques boost the retrieval performance
- We discover that the orientation information between cross-view images can be estimated when the alignment is unknown

More on visual explanation

- Zhu, Sijie, Taojiannan Yang, and Chen Chen. "Visual explanation for deep metric learning." IEEE Transactions on Image Processing (2021): 7593-7607.



Outline

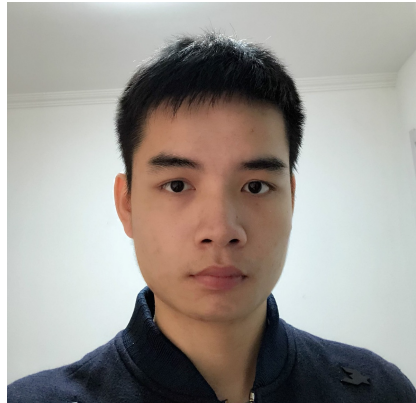
- Introduction (image geo-localization)
- **Cross-view image geo-localization**
 - **Orientational alignment in image geo-localization**

Sijie Zhu, Taojiannan Yang, Chen Chen, "Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation" Winter Conference on Applications of Computer Vision (WACV), 2021.
 - **Spatial alignment in image geo-localization**

Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
 - **Vision transformer for image geo-localization**

Zhu, Sijie, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Future work

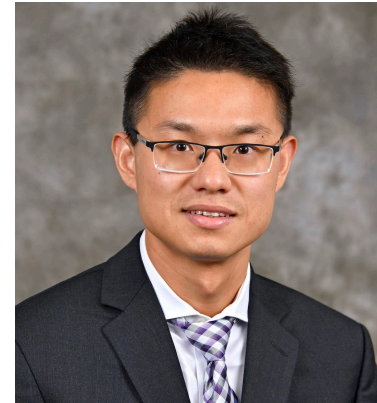
VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval



Sijie Zhu



Taojiannan Yang



Chen Chen

Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021

Spatial alignment

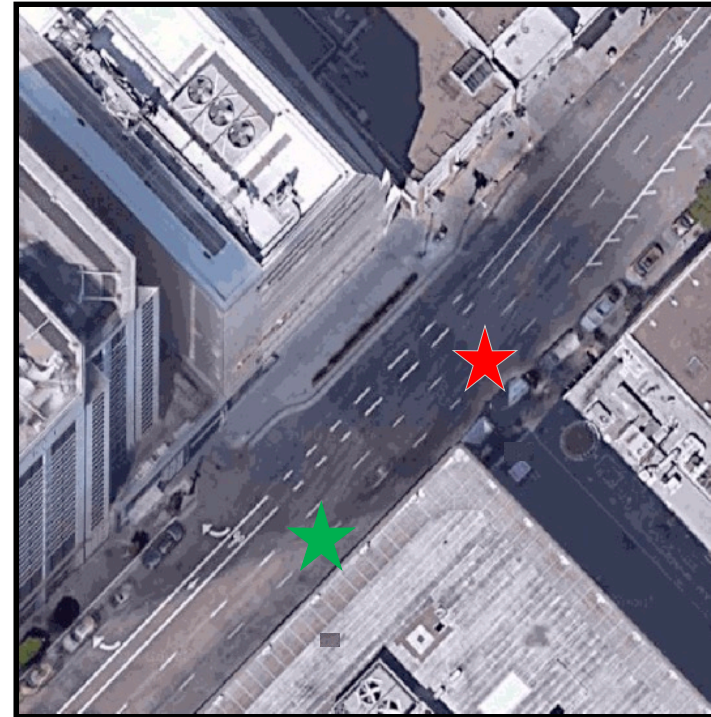
- Existing works simply assume that each query ground-view image has one corresponding reference aerial-view image whose **center is exactly aligned at the location of the query image**.
- This is not practical for real-world applications, because the query image may be generated at arbitrary locations in the area of interest and the reference images should be captured before the queries emerge.

Spatial alignment

Query

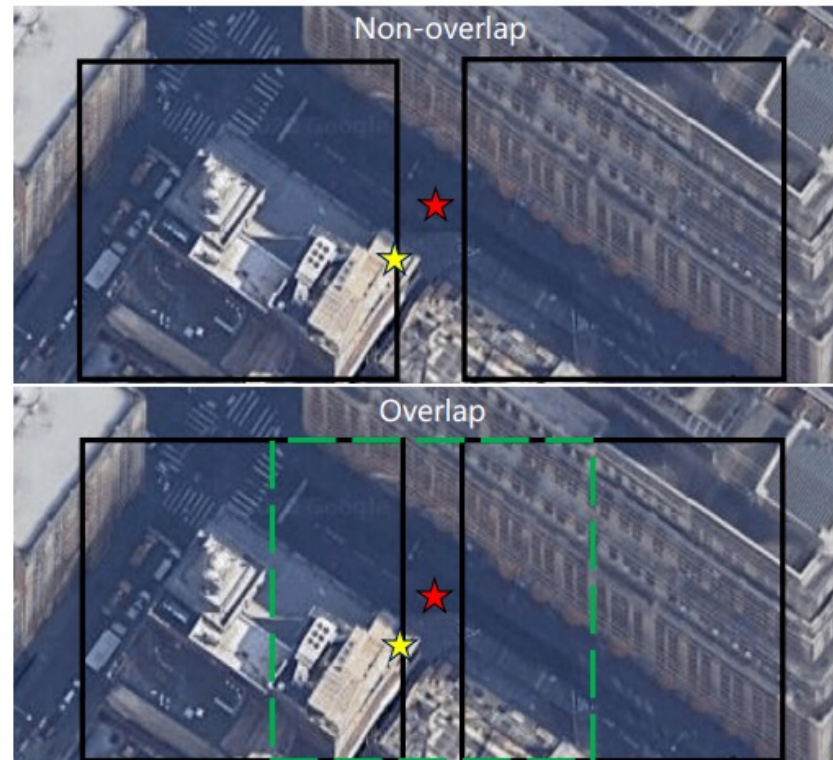


Reference



VIGOR dataset

Dataset Setting: given an area of interest (AOI), the reference aerial images are densely sampled to achieve a seamless coverage of the AOI and the street-view queries are captured at arbitrary locations.

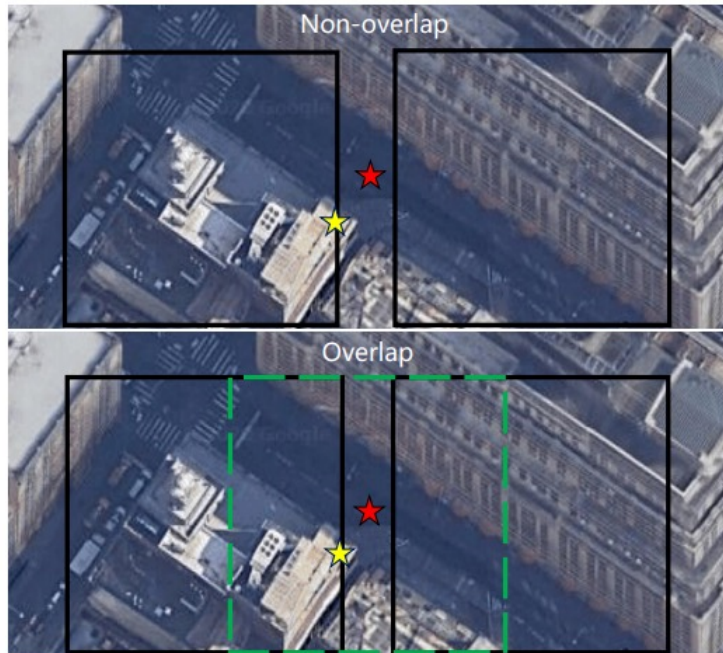


(a) Non-overlap vs Overlap Sampling.

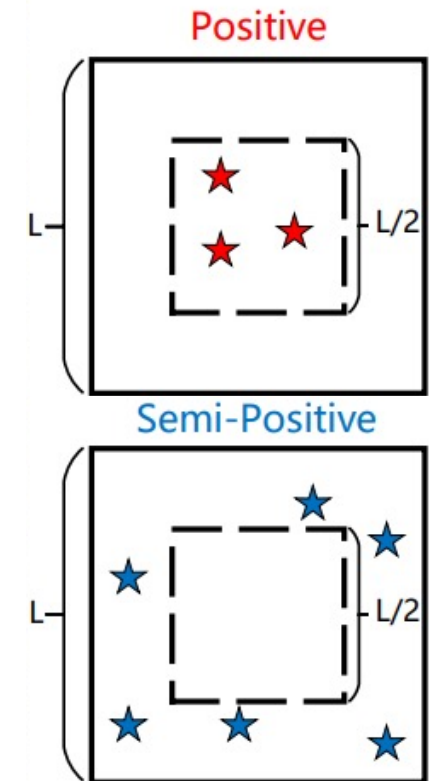
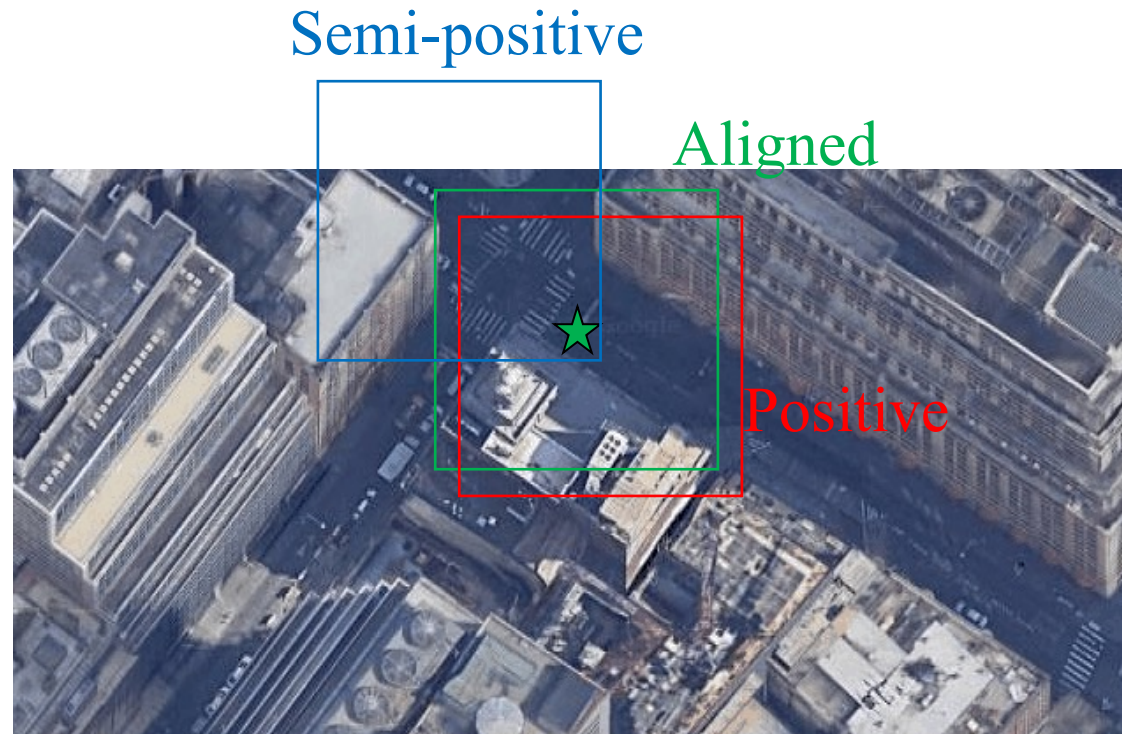
(b) I

VIGOR dataset

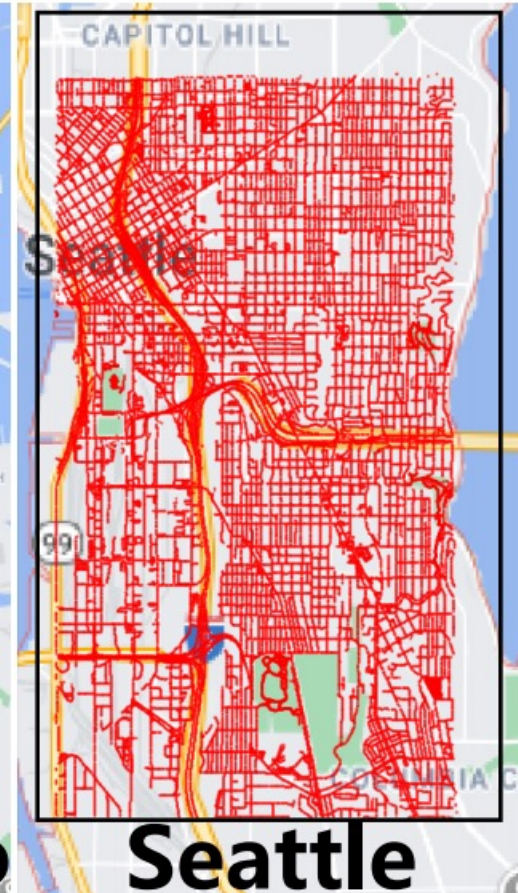
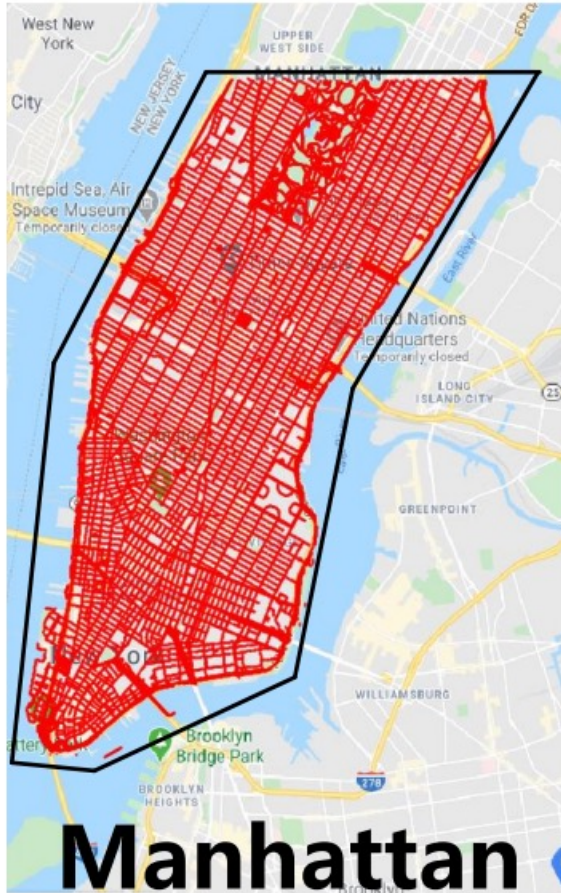
Dataset Setting: beyond one-to-one correspondence (one to many)



(a) Non-overlap vs Overlap Sampling. (b) I



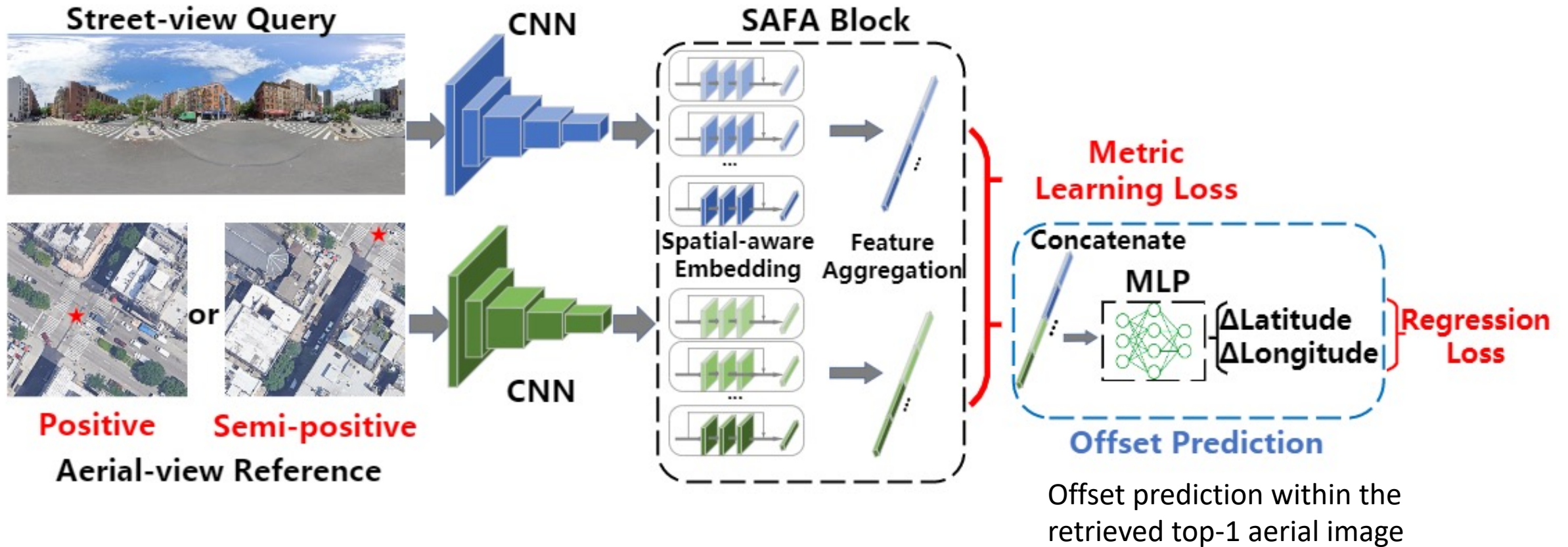
Data Distribution



Datasets Comparison

	Vo [24]	CVACT [11]	CVUSA [27]	VIGOR (proposed)
Satellite images	~ 450,000	128,334	44,416	90,618
Panoramas in total	~ 450,000	128,334	44,416	238,696
Panoramas after balancing	-	-	-	105,214
Street-view GPS locations	Aligned	Aligned	Aligned	Arbitrary
Full panorama	✗	✓	✓	✓
Multiple cities	✓	✗	✓	✓
Orientation information	✓	✓	✓	✓
Evaluation in terms of meters	✗	✗	✗	✓
Seamless coverage on area of interest	✗	✗	✗	✓
Number of references covering each query	1	1	1	4

Coarse-to-fine Cross-view Localization



Beyond One-to-one

How to make use of the semi-positive images?

Directly considering semi-positive as positive results in a low accuracy.

We force the ratio of the similarities in the embedding space to be close to the ratio of IOUs.

IOU-based semi-positive assignment loss

$$\mathcal{L}_{IOU} = \left(\frac{S_{semi}}{S_{pos}} - \frac{IOU_{semi}}{IOU_{pos}} \right)^2$$

Semi-positive Assignment	Same-Area				Cross-Area			
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
No semi-positive (<i>i.e.</i> baseline, $\mathcal{L}_{triplet}$)	38.0	62.9	97.6	41.8	9.2	21.1	77.8	9.9
Positive ($\mathcal{L}_{triplet}$)	20.3	45.7	97.9	25.4	2.7	7.6	58.2	3.1
IOU ($\mathcal{L}_{triplet} + \mathcal{L}_{IOU}$)	41.1	65.9	98.3	44.8	10.7	23.5	79.3	11.4

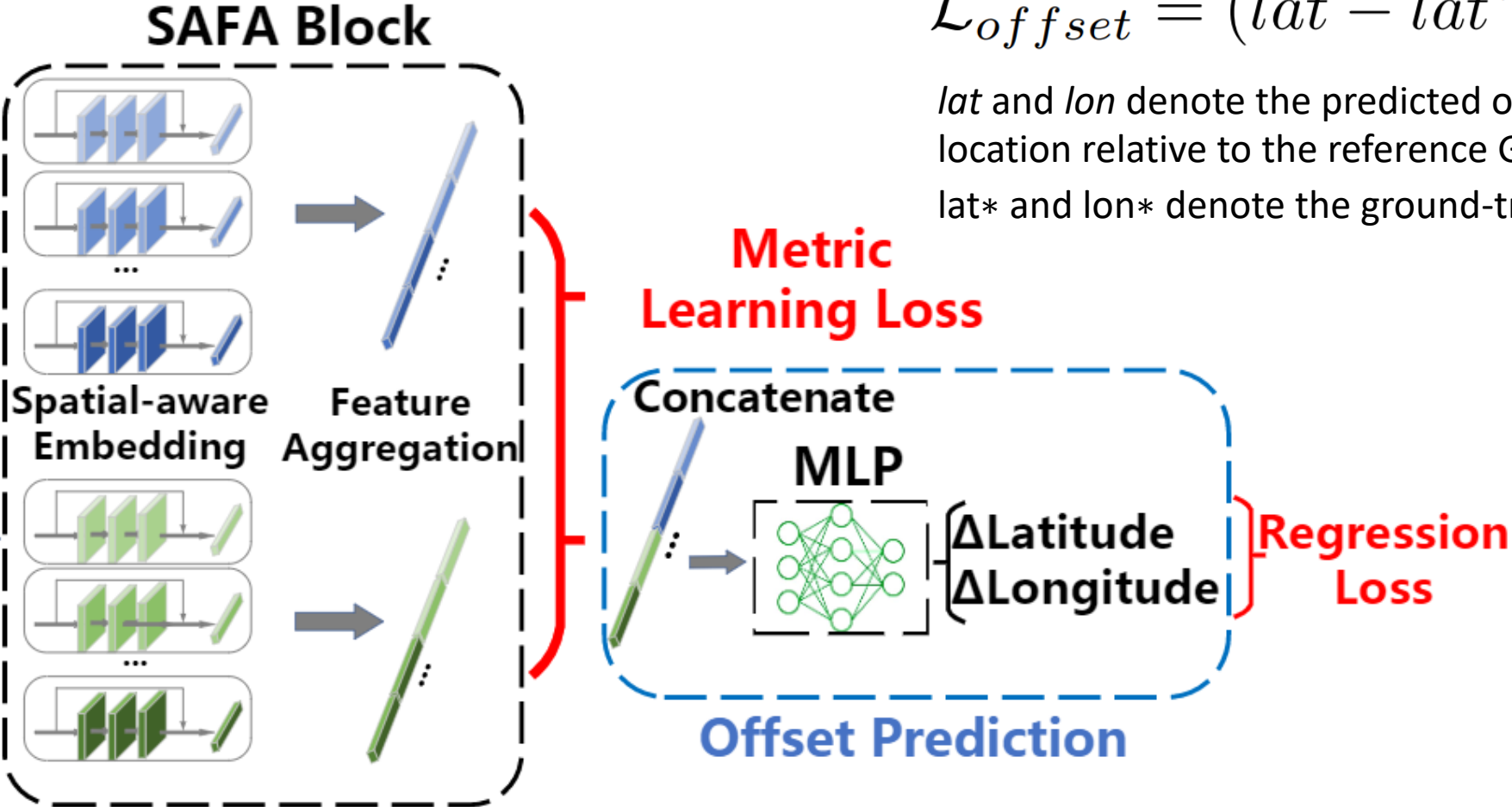
Beyond Retrieval

Offset prediction within the retrieved top-1 aerial image

$$\mathcal{L}_{offset} = (lat - lat^*)^2 + (lon - lon^*)^2$$

lat and *lon* denote the predicted offset of the query GPS location relative to the reference GPS

*lat** and *lon** denote the ground-truth offset



Comparison with State-of-the-art

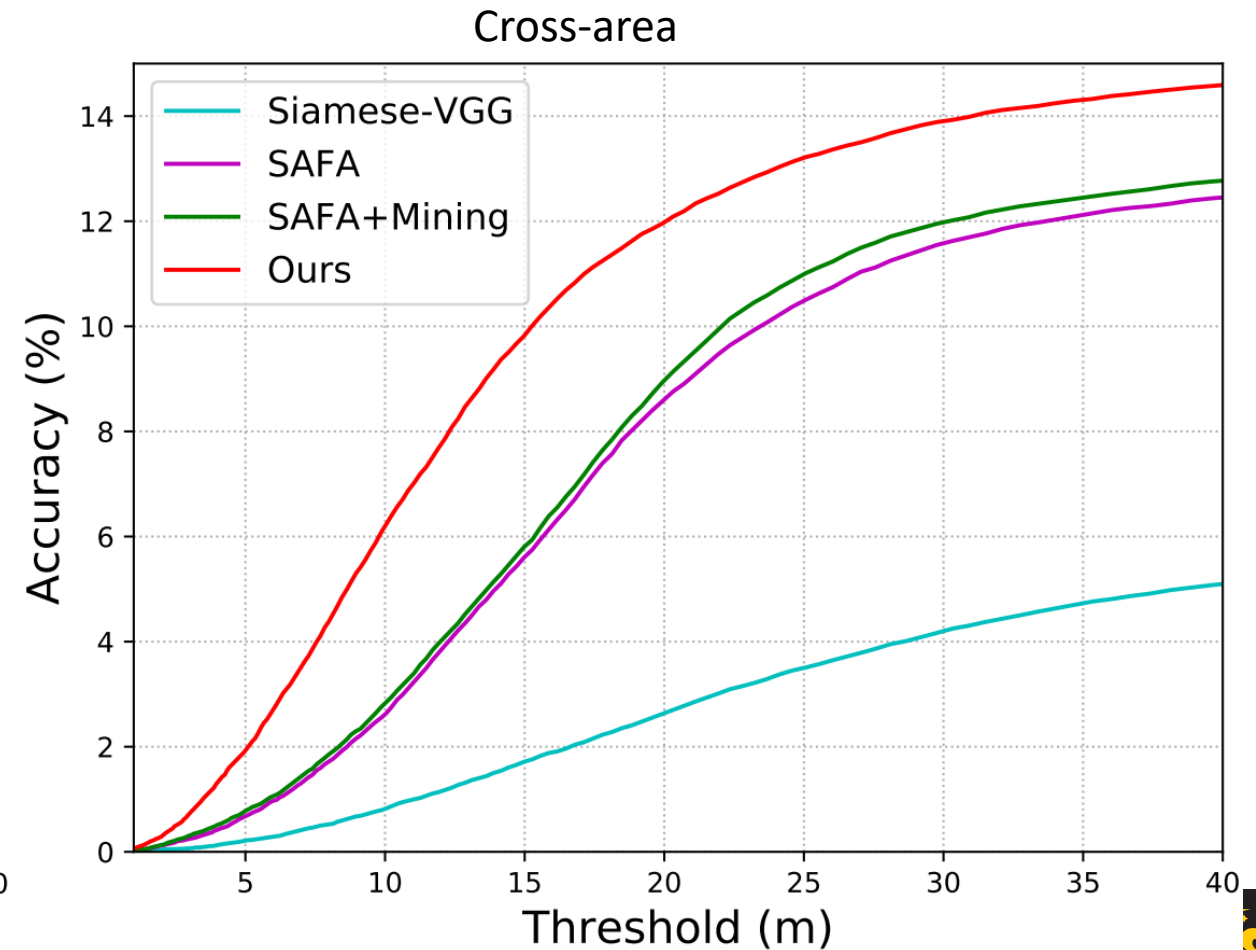
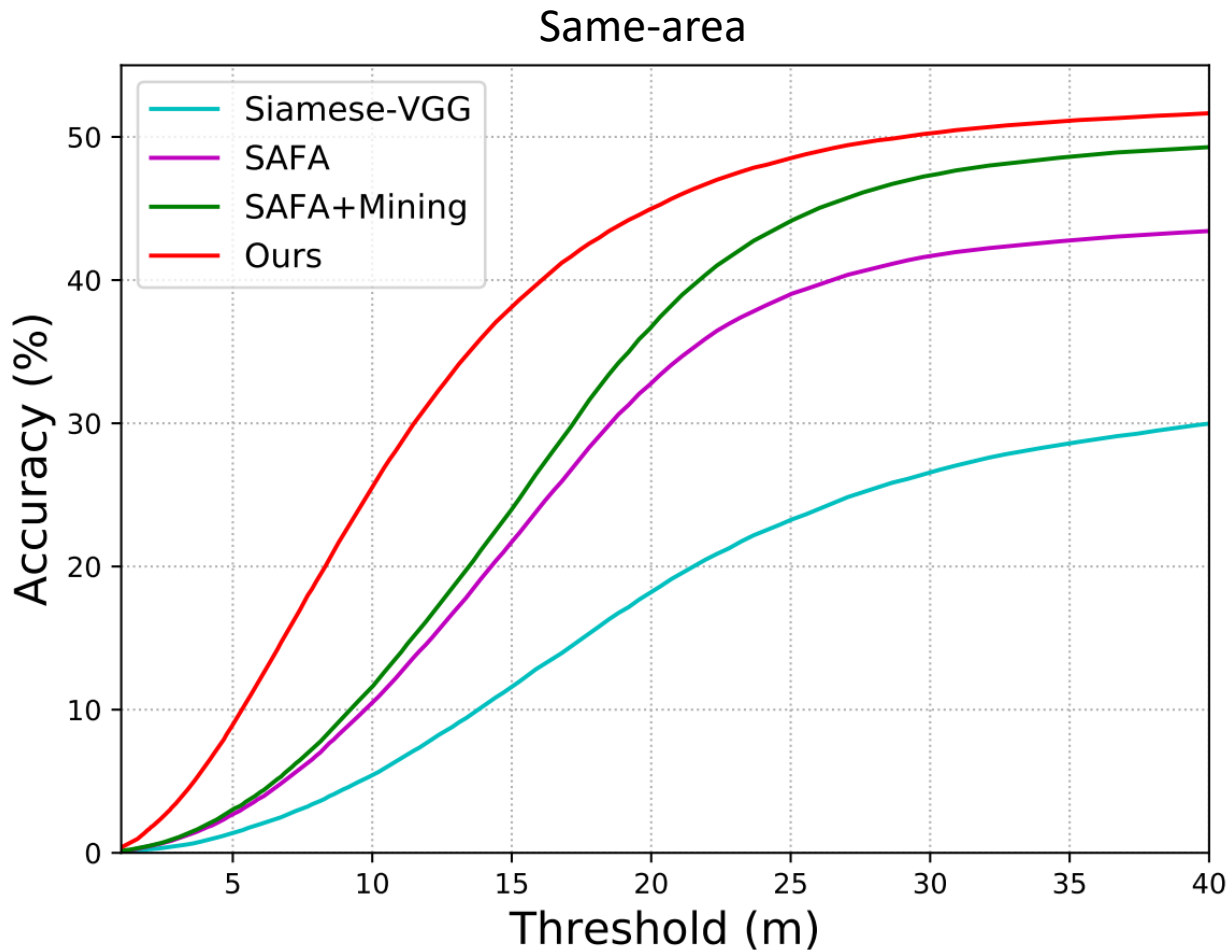
- Retrieval Performance

	Same-Area				Cross-Area		
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%
Siamese-VGG ($\mathcal{L}_{triplet}$)	18.1	42.5	97.5	21.2	2.7	8.2	61.7
SAFA ($\mathcal{L}_{triplet}$)	33.9	58.4	98.2	36.9	8.2	19.6	77.6
SAFA+Mining (baseline, $\mathcal{L}_{triplet}$)	38.0	62.9	97.6	41.8	9.2	21.1	77.8
Ours (\mathcal{L}_{hybrid})	41.1	65.8	98.4	44.7	11.0	23.6	80.2

$$\mathcal{L}_{hybrid} = \mathcal{L}_{triplet} + \mathcal{L}_{IOU} + \mathcal{L}_{offset}$$

Comparison with State-of-the-art

- Localization in terms of meters



The Effect of Offset Prediction

- Localization in terms of meters

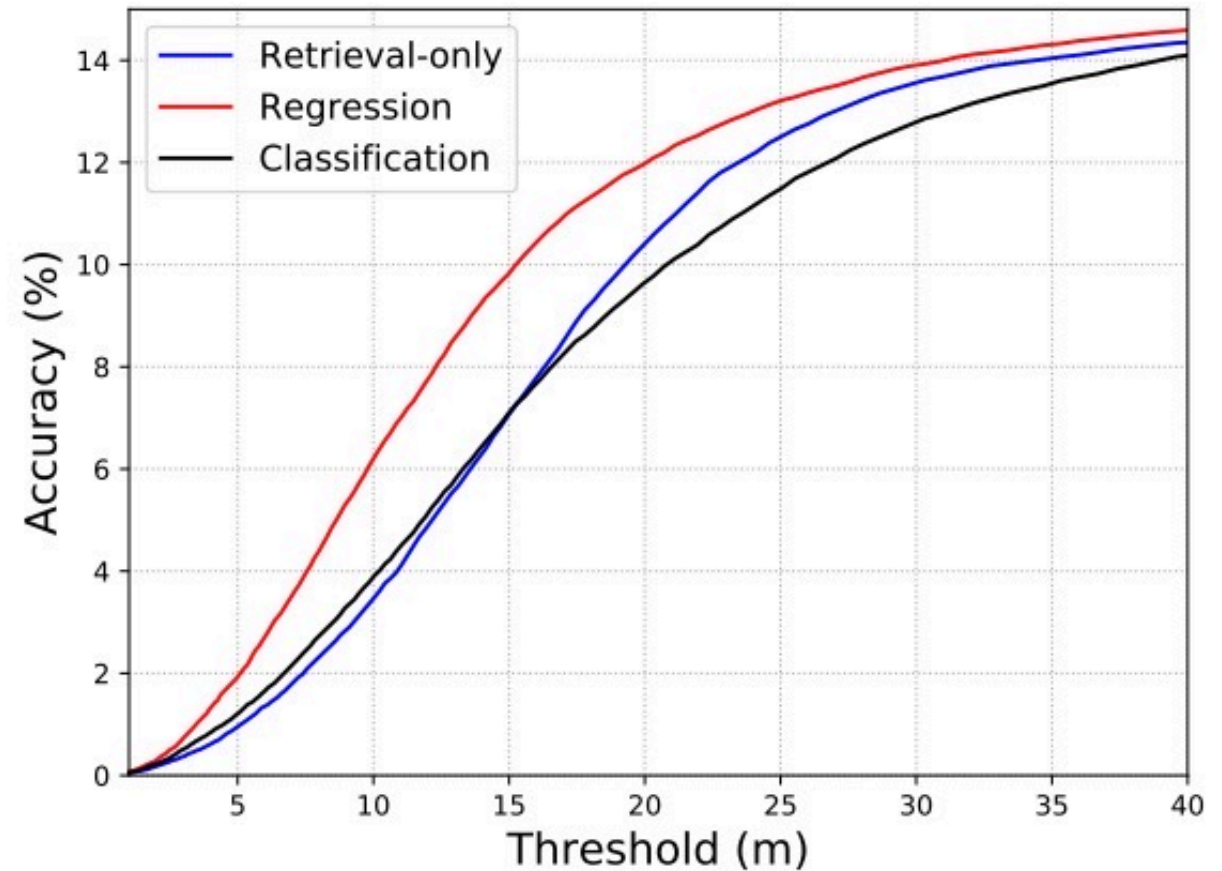
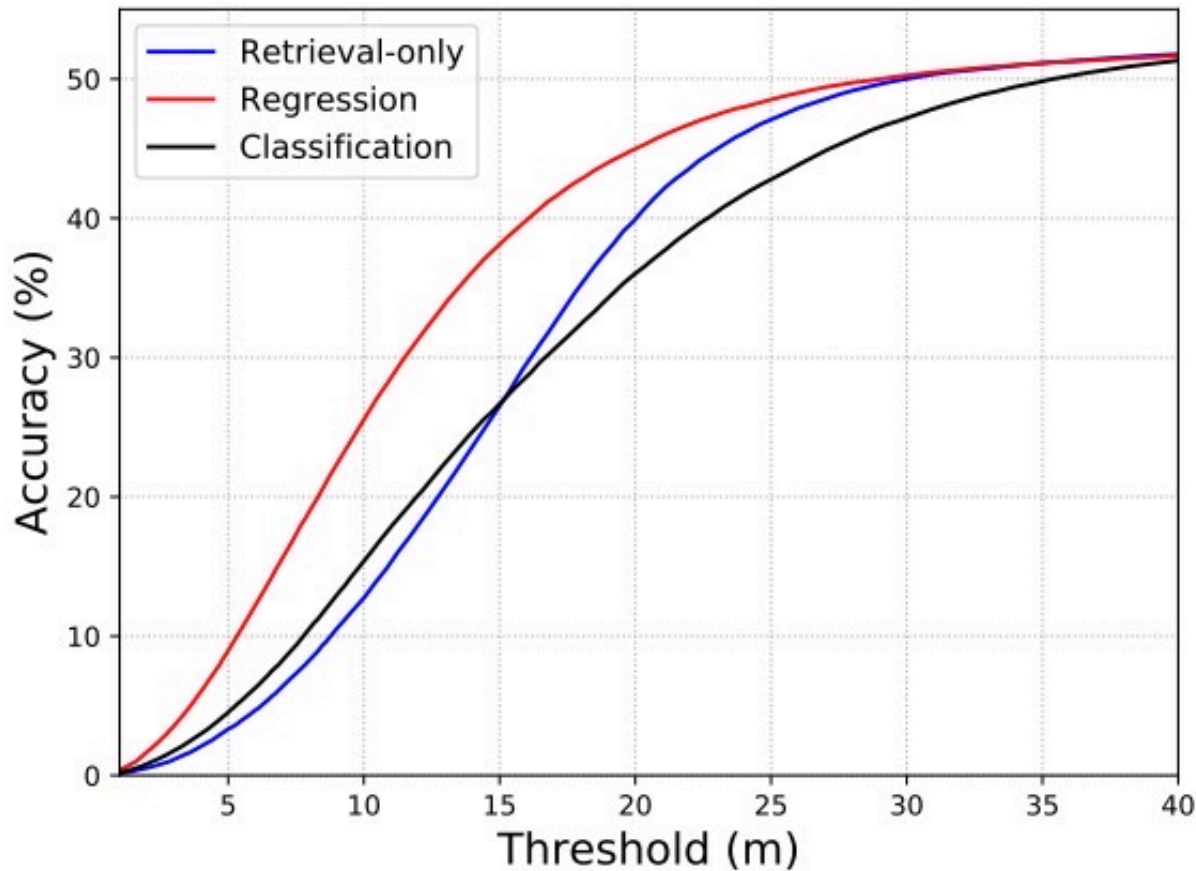
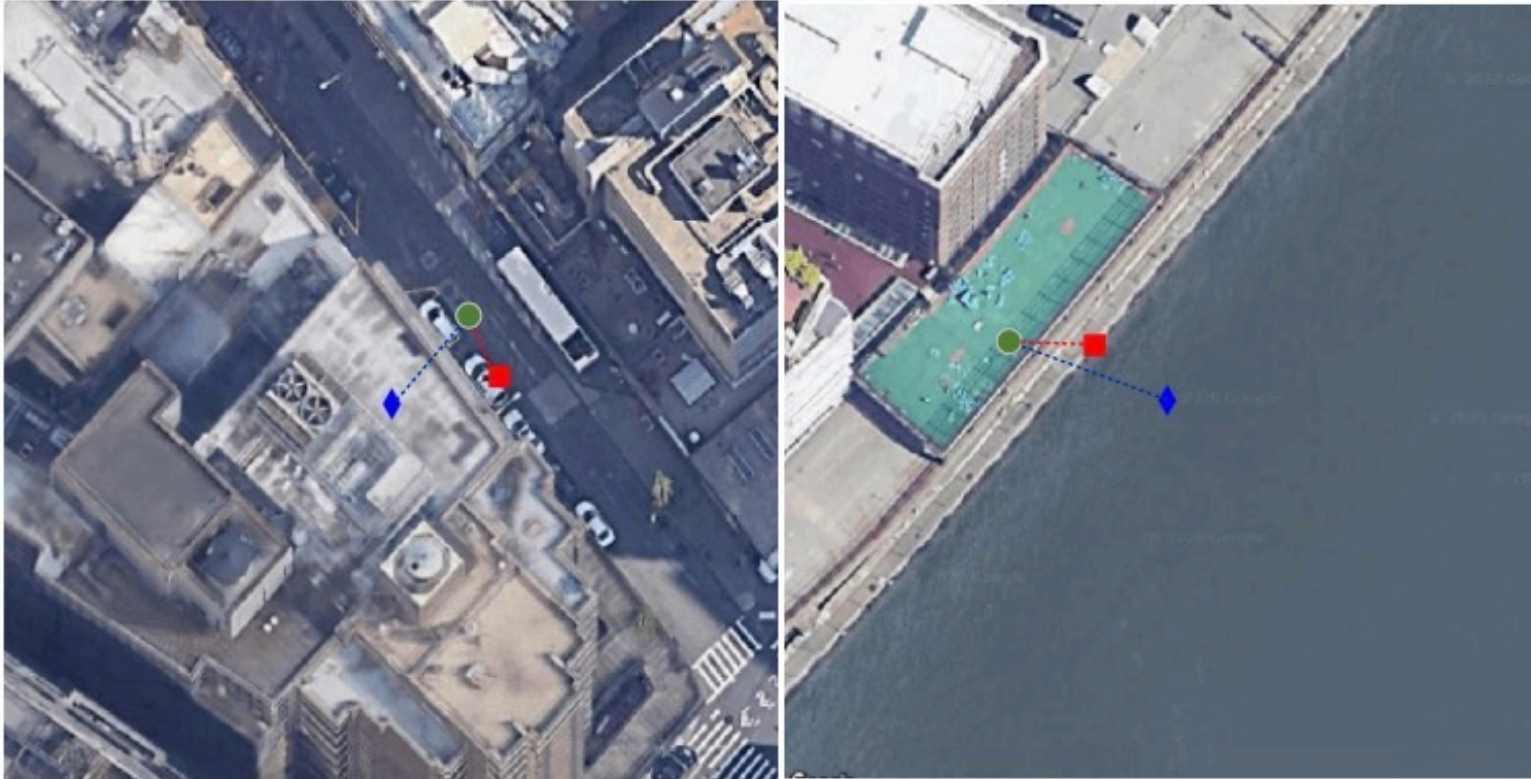


Figure 6. Same-area (left) and cross-area (right) meter-level localization accuracy of different offset prediction methods.

The Effect of Offset Prediction



- Ground truth location
- ◆ No offset prediction
- With offset prediction

Figure 8. Case study on meter-level refinement within the retrieved aerial image. Red square, green circle and blue diamond denote the final prediction with regression, ground-truth, and center (*i.e.* the prediction with only retrieval), respectively.

Noisy GPS Refinement

- Retrieval in a searching scope

Search Scope	Same-Area		Cross-Area	
	Top-1	Top-5	Top-1	Top-5
All	41.1	65.8	11.0	23.6
1000 <i>m</i>	49.2	76.7	19.9	41.5
500 <i>m</i>	54.1	82.6	26.4	53.3
200 <i>m</i>	60.9	90.6	37.7	72.0

Summary

- We propose a new benchmark for cross-view image geo-localization beyond one-to-one retrieval, which is a more realistic setting for real-world applications.
- The proposed method significantly Improves 10-meter-level accuracy:
 - 11.4% → 25.5% for same-area evaluation
 - 2.8% → 6.2% for cross-area evaluation
- Code and dataset are available at <https://github.com/Jeff-Zilence/VIGOR>

Outline

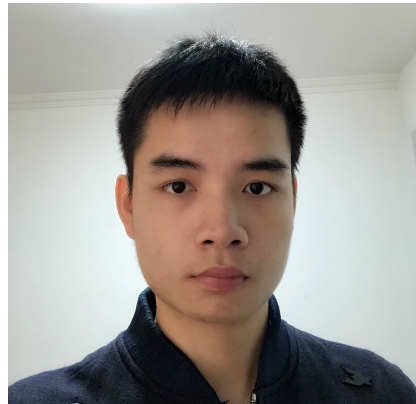
- Introduction (image geo-localization)
- **Cross-view image geo-localization**
 - **Orientational alignment in image geo-localization**

Sijie Zhu, Taojiannan Yang, Chen Chen, "Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation" Winter Conference on Applications of Computer Vision (WACV), 2021.
 - **Spatial alignment in image geo-localization**

Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
 - **Vision transformer for image geo-localization**

Zhu, Sijie, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Future work

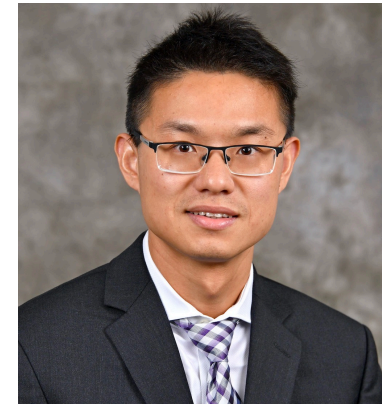
TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization



Sijie Zhu



Mubarak Shah

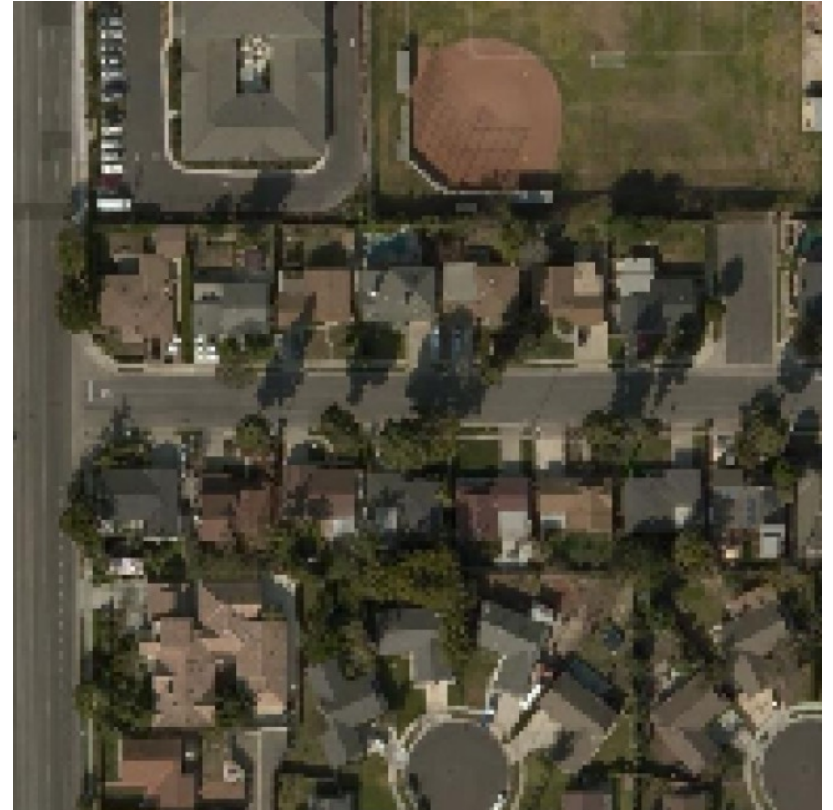


Chen Chen

Zhu, Sijie, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Cross-view Image geo-localization

Domain Gap



Predominant CNN-based Methods on CVUSA

Method	R@1	R@5	R@10	R@1%
Workman [30]	-	-	-	34.30
Zhai [34]	-	-	-	43.20
CVM-Net [10]	22.47	49.98	63.18	93.62
Liu [14]	40.79	66.82	76.36	96.12
Reweight [3]	-	-	-	98.30
Regmi [19]	48.75	-	81.27	95.98
Revisit [35]	70.40	-	-	99.10
SAFA [21]	81.15	94.23	96.85	99.49
†SAFA [21]	89.84	96.93	98.14	99.64
†Shi [22]	91.96	97.50	98.54	99.67
†Toker [26]	92.56	97.55	98.33	99.57

Suffer from
Domain Gap

Polar Transform

Polar Transform works well on CVUSA dataset



Aerial



Ground



Polar-transformed Aerial Image

$$x_i^s = \frac{A_a}{2} + \frac{A_a}{2} \frac{y_i^t}{H_g} \sin\left(\frac{2\pi}{W_g} x_i^t\right)$$
$$y_i^s = \frac{A_a}{2} - \frac{A_a}{2} \frac{y_i^t}{H_g} \cos\left(\frac{2\pi}{W_g} x_i^t\right)$$

Shi, Yujiao, et al. "Spatial-aware feature aggregation for image based cross-view geo-localization." Advances in Neural Information Processing Systems 32 (2019).

Polar Transform doesn't Work on VIGOR dataset

Aerial-view Reference



Street-view Query



Polar Transform

VIGOR

SAFA [21]	33.93
SAFA+Polar	24.13

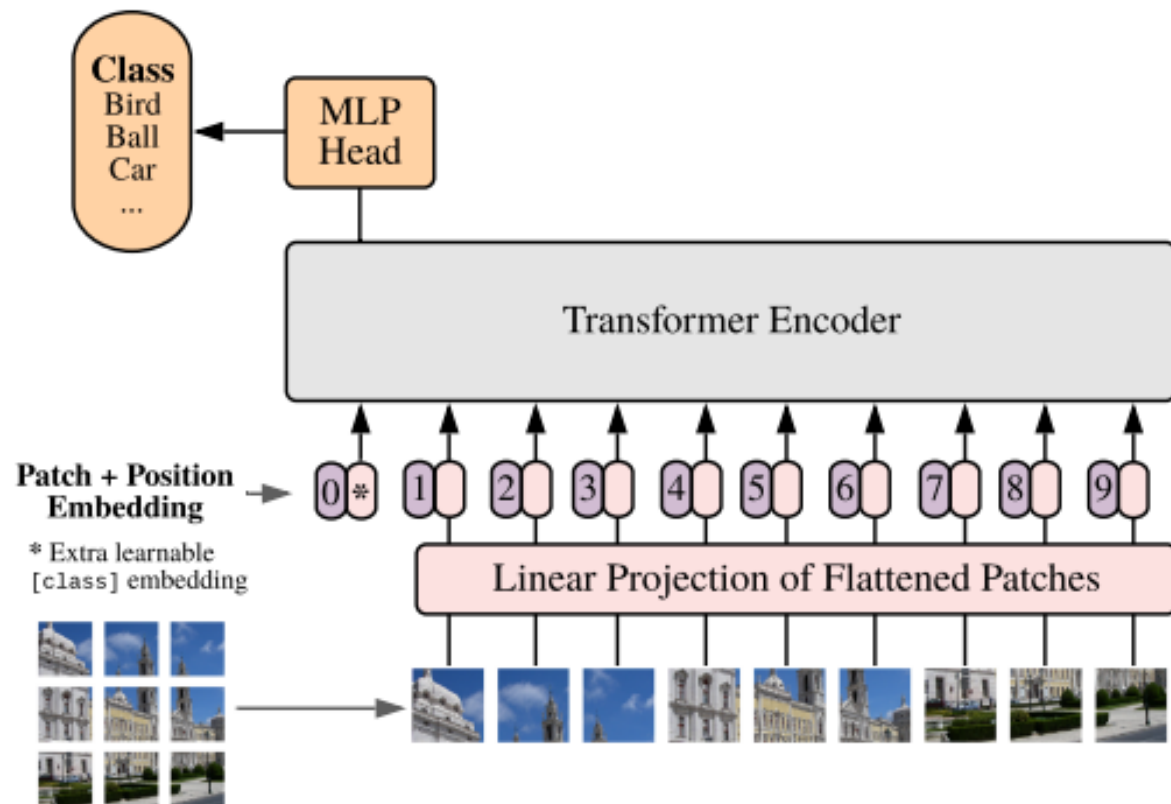


How to Bridge the Domain Gap?

✗ Polar Transform + CNN

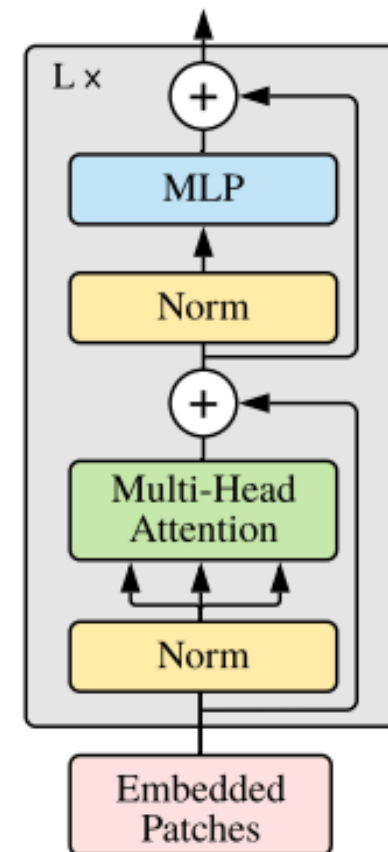
✓ Vision transformer

Vision Transformer (ViT)



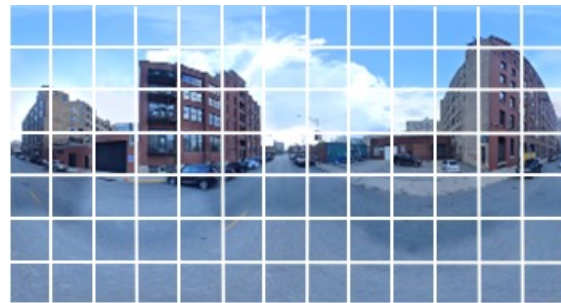
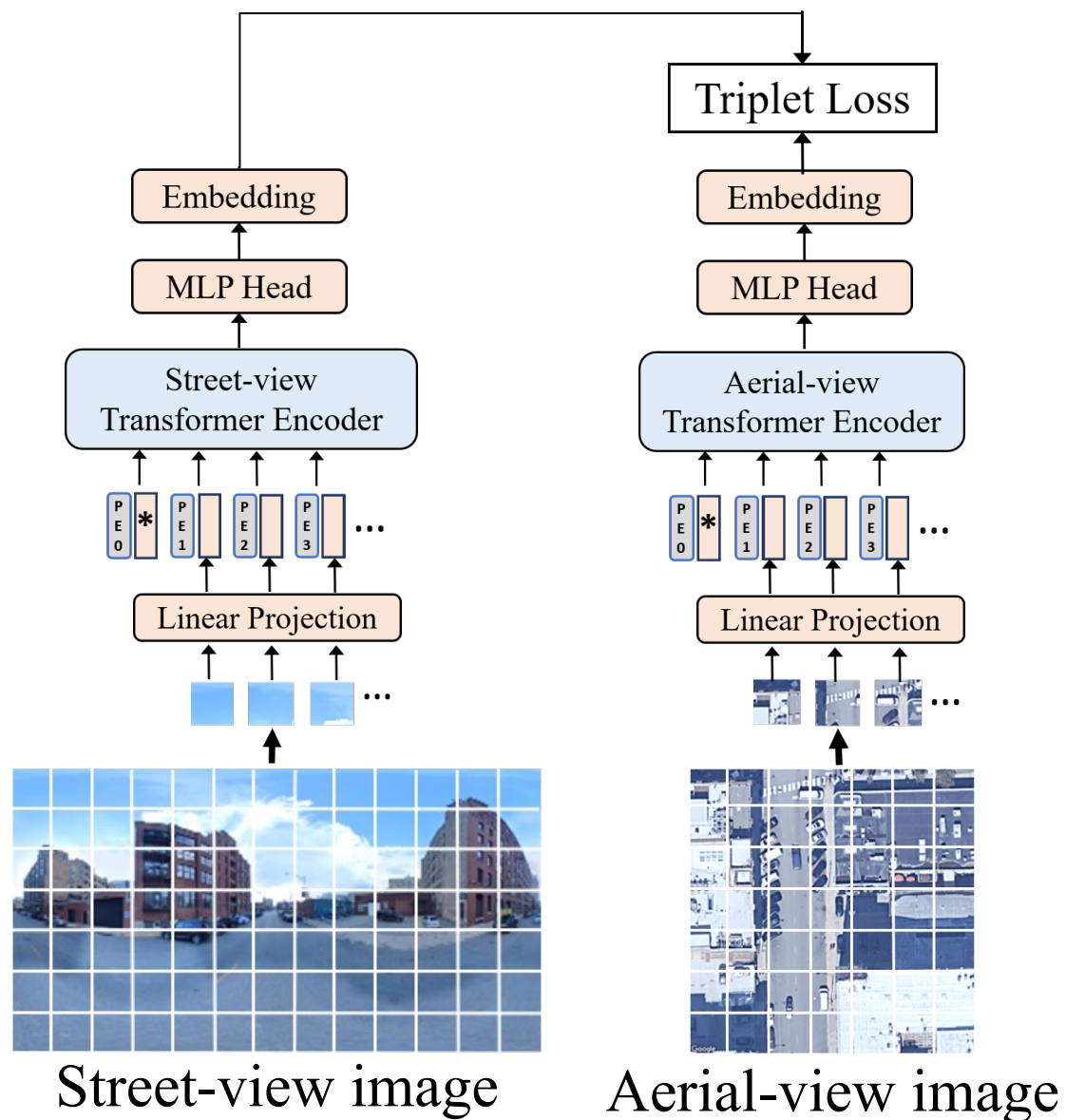
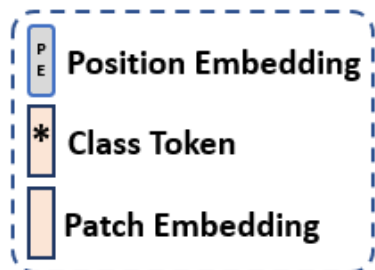
- Explicit positional information
- Global attention

Transformer Encoder



TransGeo

Stage 1

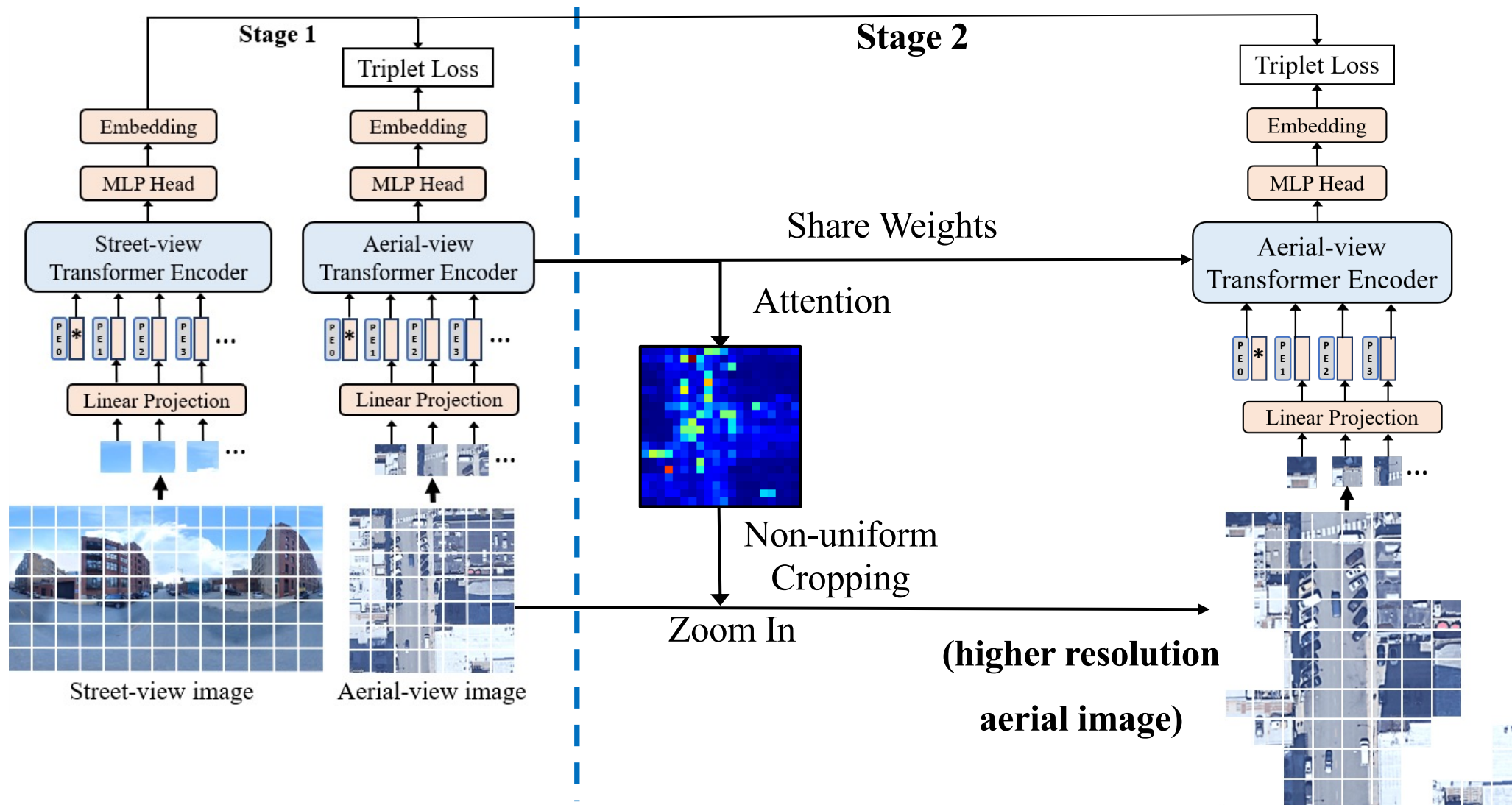


Street-view image

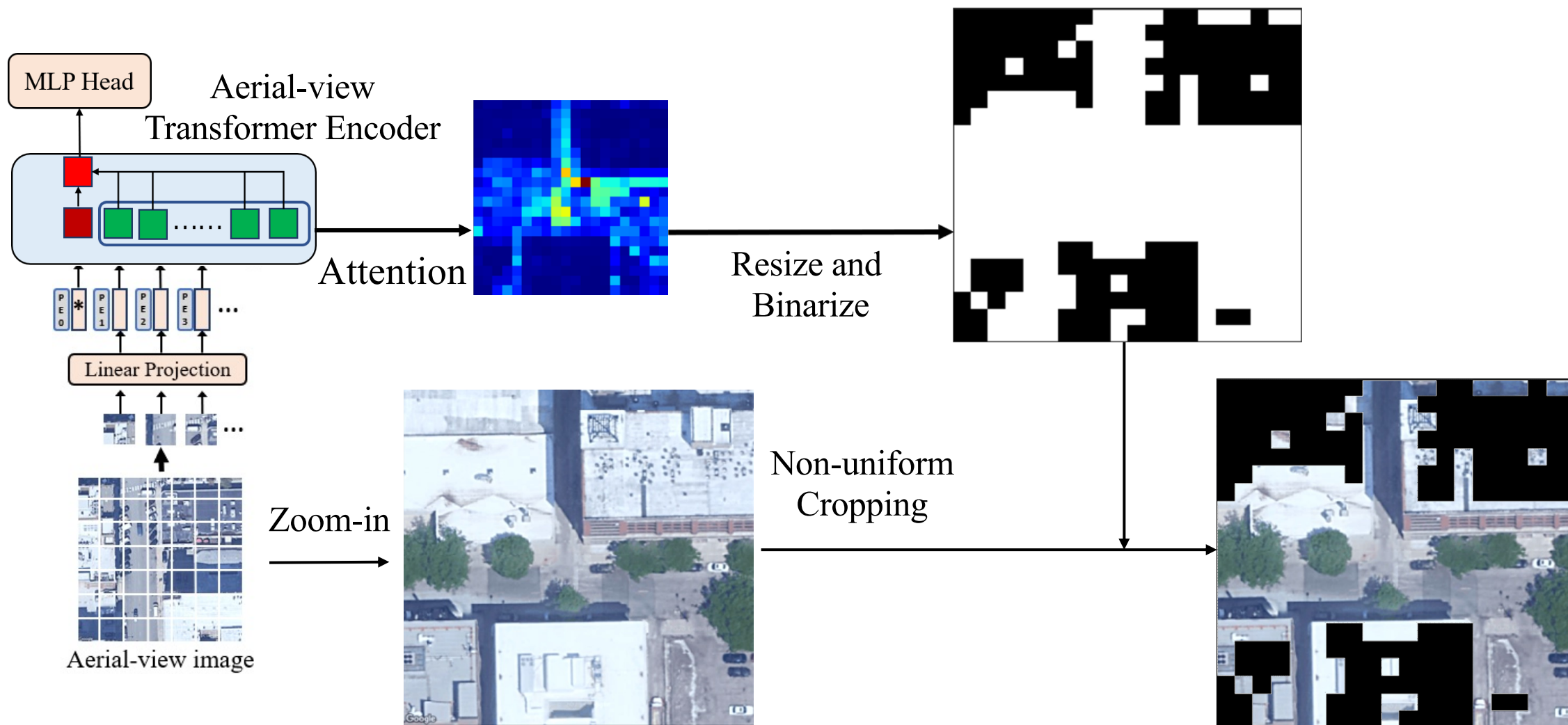


Aerial-view image

Stage 2 - Attend and Zoom-in



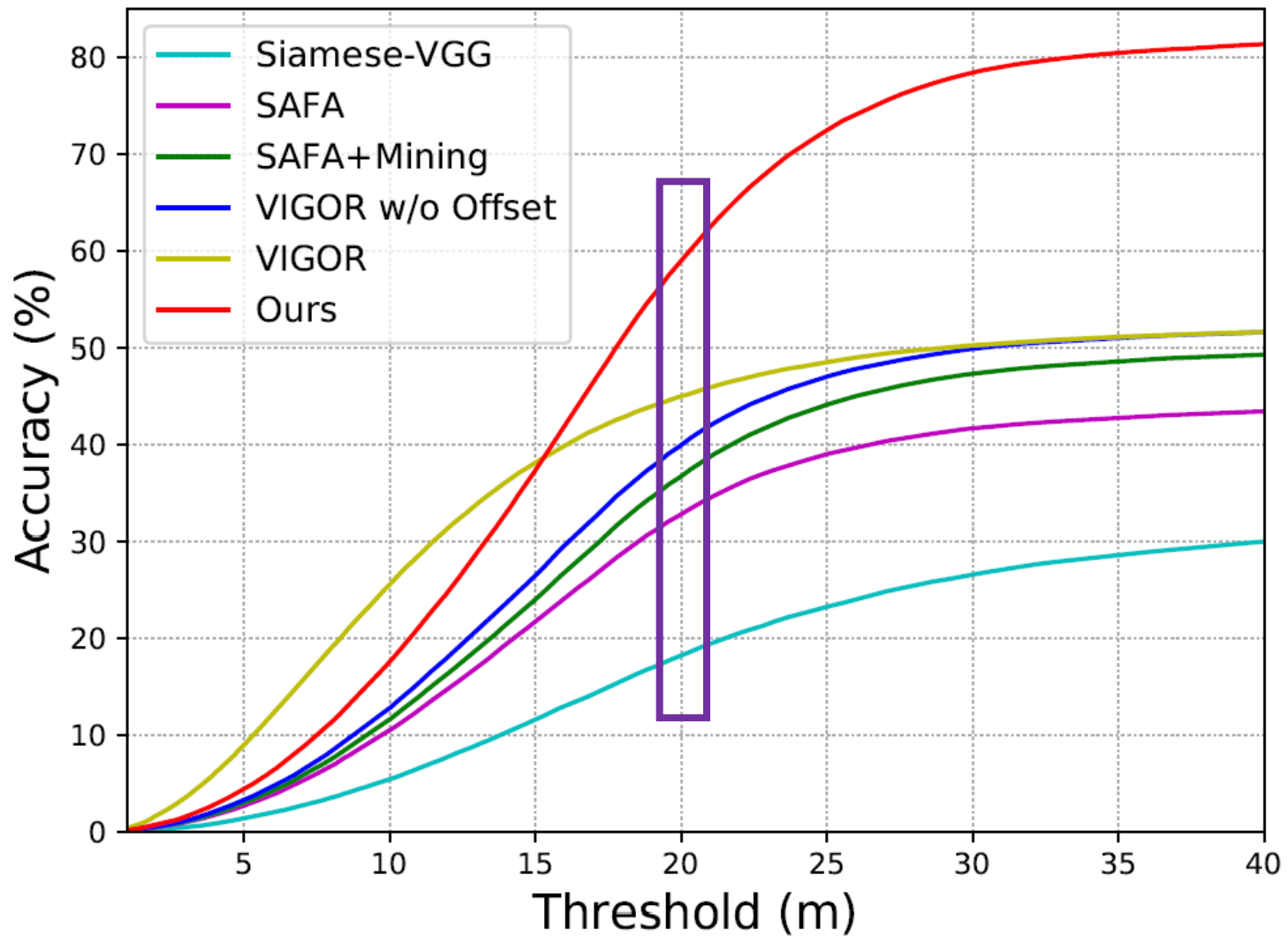
Non-uniform Cropping



Retrieval Performance on VIGOR

	R@1	R@5	R@10	R@1%	Hit
Siamese-VGG [35]	18.69	43.64	55.36	97.55	21.90
SAFA [21]	33.93	58.42	68.12	98.24	36.87
SAFA+Mining [36]	38.02	62.87	71.12	97.63	41.81
VIGOR [36]	41.07	65.81	74.05	98.37	44.71
Ours	61.48	87.54	91.88	99.56	73.09

Meter-level Evaluation



Performance on CVUSA

- Query is exactly at the center of reference aerial image.

Method	R@1	R@5	R@10	R@1%
Workman [30]	-	-	-	34.30
Zhai [34]	-	-	-	43.20
CVM-Net [10]	22.47	49.98	63.18	93.62
Liu [14]	40.79	66.82	76.36	96.12
Reweight [3]	-	-	-	98.30
Regmi [19]	48.75	-	81.27	95.98
Revisit [35]	70.40	-	-	99.10
SAFA [21]	81.15	94.23	96.85	99.49
EgoTR [31] (arXiv)	91.99	97.68	98.65	99.75
†SAFA [21]	89.84	96.93	98.14	99.64
†Shi [22]	91.96	97.50	98.54	99.67
†Toker [26]	92.56	97.55	98.33	99.57
†EgoTR [31] (arXiv)	94.05	98.27	98.99	99.67
Ours	94.08	98.36	99.04	99.77

† means using Polar Transform

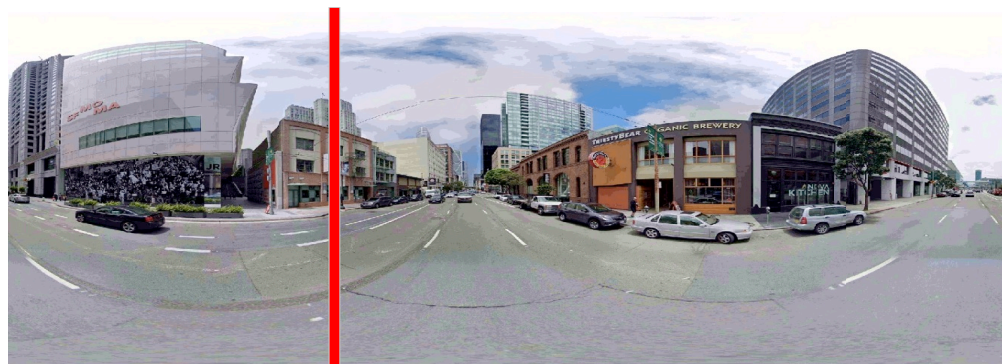
The First to Quantitatively Measure Efficiency

Method	GFLOPs	GPU Memory	Inference Time per Batch	R@1
†SAFA	42.24	10.82 GB	111 ms	89.84
Ours	11.32	9.85 GB	99 ms	94.08

Unknown Orientation



↑ North



	Same-Area				Cross-Area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
VIGOR [36]	19.10	42.13	-	95.12	1.41	4.52	-	44.60
TransGeo	47.69	79.77	86.36	99.29	5.54	14.22	19.63	66.93

Limited Field of View (FoV)

$FoV = 360^\circ$



$FoV = 180^\circ$



$FoV = 90^\circ$



	$FoV = 180^\circ$				$FoV = 90^\circ$			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
DSM [22]	48.53	68.47	75.63	93.02	16.19	31.44	39.85	71.13
TransGeo	58.22	81.33	87.66	98.13	30.12	54.18	63.96	89.18

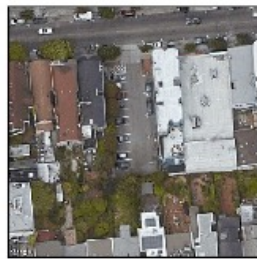
[22] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4064–4072, 2020

Qualitative Results-VIGOR

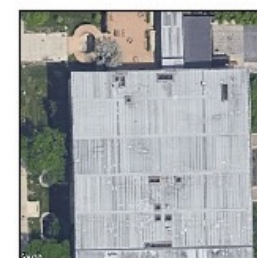
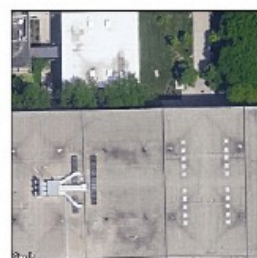
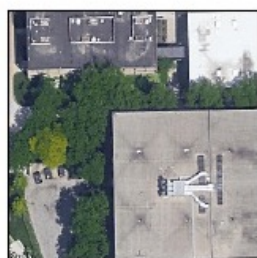
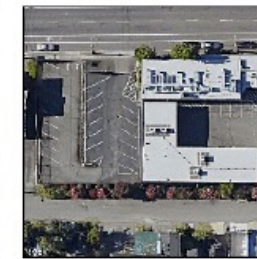
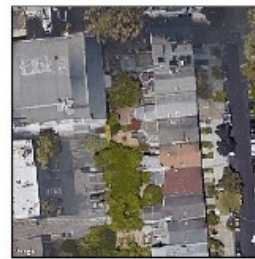
Street-view Query



Ground-truth



Retrieved Reference Images



Qualitative Results-CVUSA

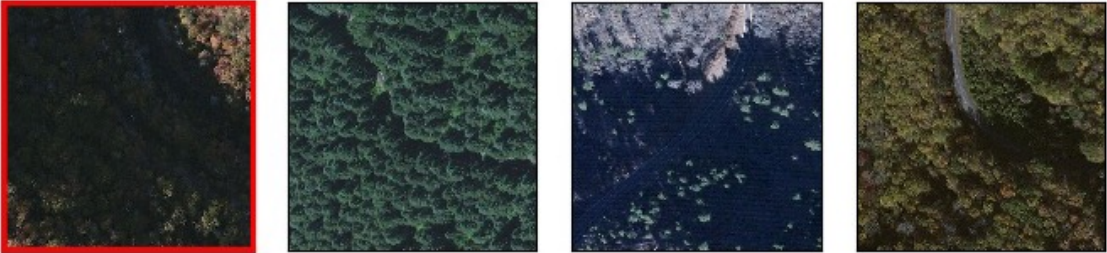
Street-view Query



Ground-truth



Retrieved Reference Images



Summary

- CNN-based methods highly rely on polar transform, but the proposed transformer-based method performs well w/o polar transform for all scenarios, due to the explicit positional embedding.
- Removing a large portion of patches from aerial view does not cause much performance drop, indicating high redundancy in this task. It could be leveraged to reduce computation or improve performance without additional cost.
- Code: <https://github.com/Jeff-Zilence/TransGeo2022>

Outline

- Introduction (image geo-localization)
- Cross-view image geo-localization
 - Orientational alignment in image geo-localization

Sijie Zhu, Taojiannan Yang, Chen Chen, "Revisiting Street-to-Aerial View Image Geo-localization and Orientation Estimation" Winter Conference on Applications of Computer Vision (WACV), 2021.
 - Spatial alignment in image geo-localization

Zhu, Sijie, Taojiannan Yang, and Chen Chen. "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
 - Vision transformer for image geo-localization

Zhu, Sijie, Mubarak Shah, and Chen Chen. "TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Future work

Future Work – Same-view + Cross-view

- Geo-localization with Multi-view Reference

Same-view



Query



Reference

Cross-view



Query



Reference

Geo-localization with Multi-view Reference

- Both street-view and aerial images exist in most cities.
- They are complimentary to each other:
 - Same-view has better performance but does not have full-coverage.
 - Cross-view is easy to collect reference but has low accuracy.
- Two separated fields can be combined.

Cross-view Video Geo-localization

- Query: ground video



Shruti Vyas, Chen Chen, Mubarak Shah. “GAMa: Cross-view Video Geo-localization”, European Conference on Computer Vision (ECCV), 2022 (a new dataset is collected)

References

- Zhu, Sijie, Taojiannan Yang, and Chen Chen. “Revisiting street-to-aerial view image geo-localization and orientation estimation.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.
- Zhu, Sijie, Taojiannan Yang, and Chen Chen. “Visual explanation for deep metric learning.” IEEE Transactions on Image Processing (TIP), 30 (2021): 7593-7607.
- Zhu, Sijie, Taojiannan Yang, and Chen Chen. “VIGOR: Cross-view image geo-localization beyond one-to-one retrieval.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.
- Zhu, Sijie, Mubarak Shah, and Chen Chen. “TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
- Shruti Vyas, Chen Chen, Mubarak Shah. “GAMa: Cross-view Video Geo-localization”, European Conference on Computer Vision (ECCV), 2022

Thank you

