Large Scale Cross View Image Geo-localization

Dr. Chen Chen

Department of Electrical and Computer Engineering

E-mail: chen.chen@uncc.edu

https://webpages.uncc.edu/cchen62/

What is image geo-localization?

Input

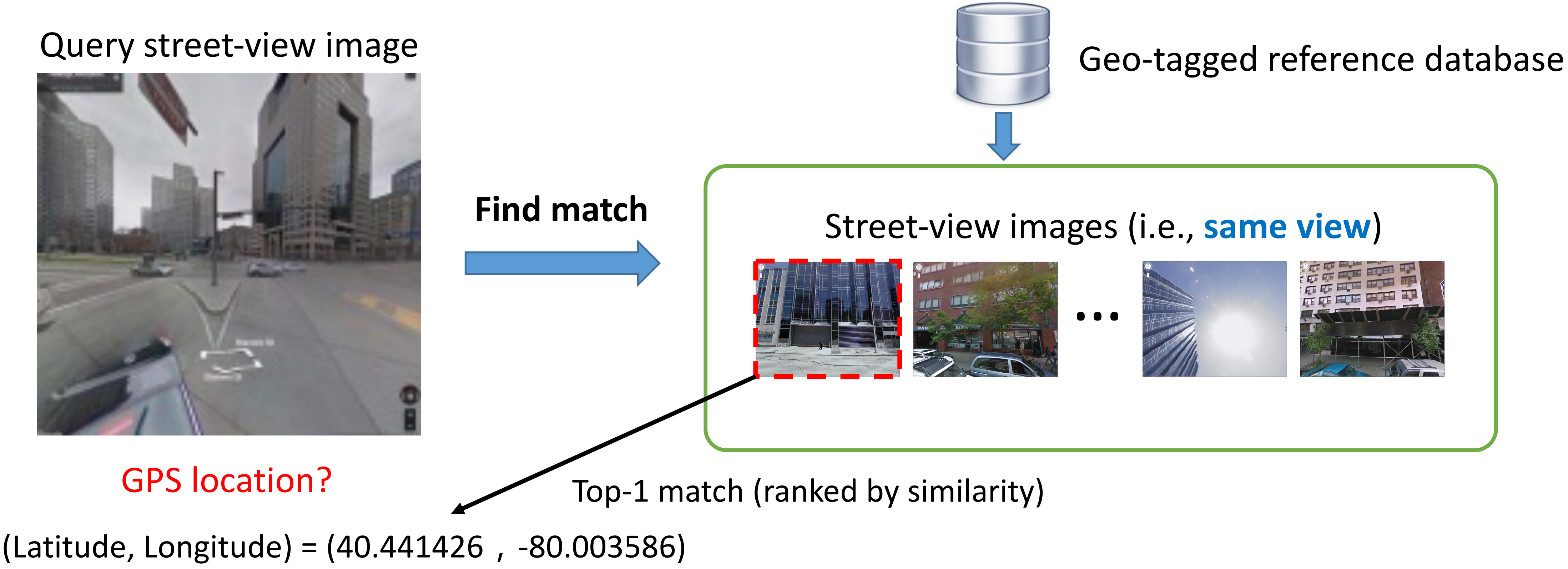
Visual Information (Images)

Output

Location in terms of Longitude and Latitude

40.4419, -79.9986

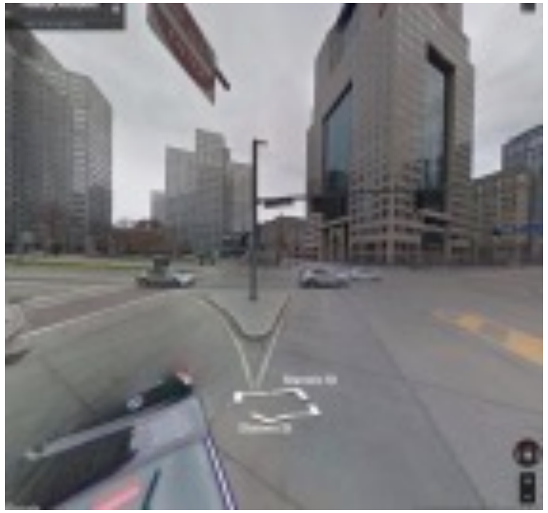
What is image geo-localization?



What is image geo-localization?

Geo-tagged reference database

Query street-view image



Find match

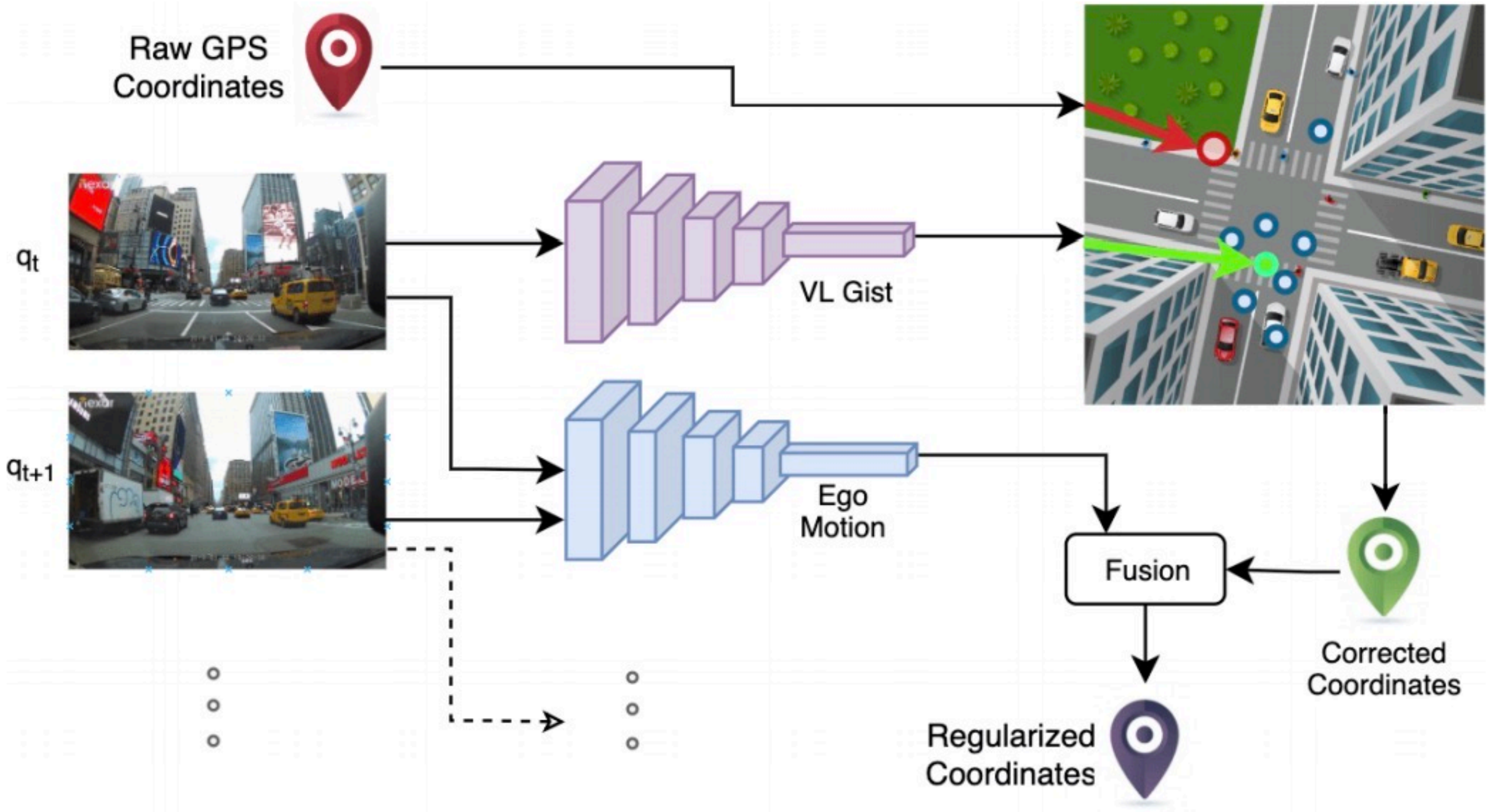
Aerial images (i.e., **cross view**)

GPS location?

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

Why is image geo-localization important?

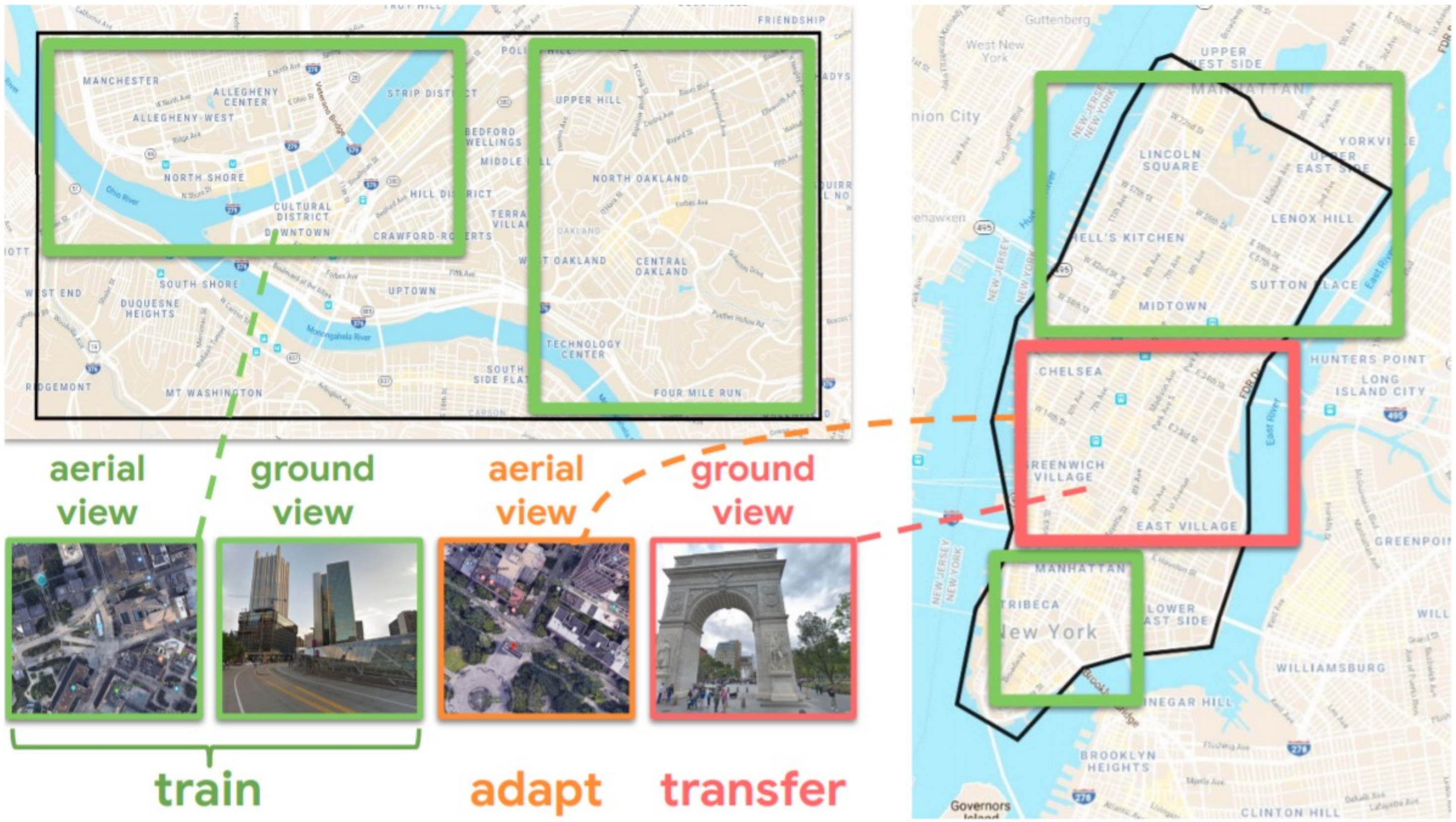
- Accurate Visual Localization for Automotive Applications



Brosh, Eli, et al. "Accurate Visual Localization for Automotive Applications." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

Why is image geo-localization important?

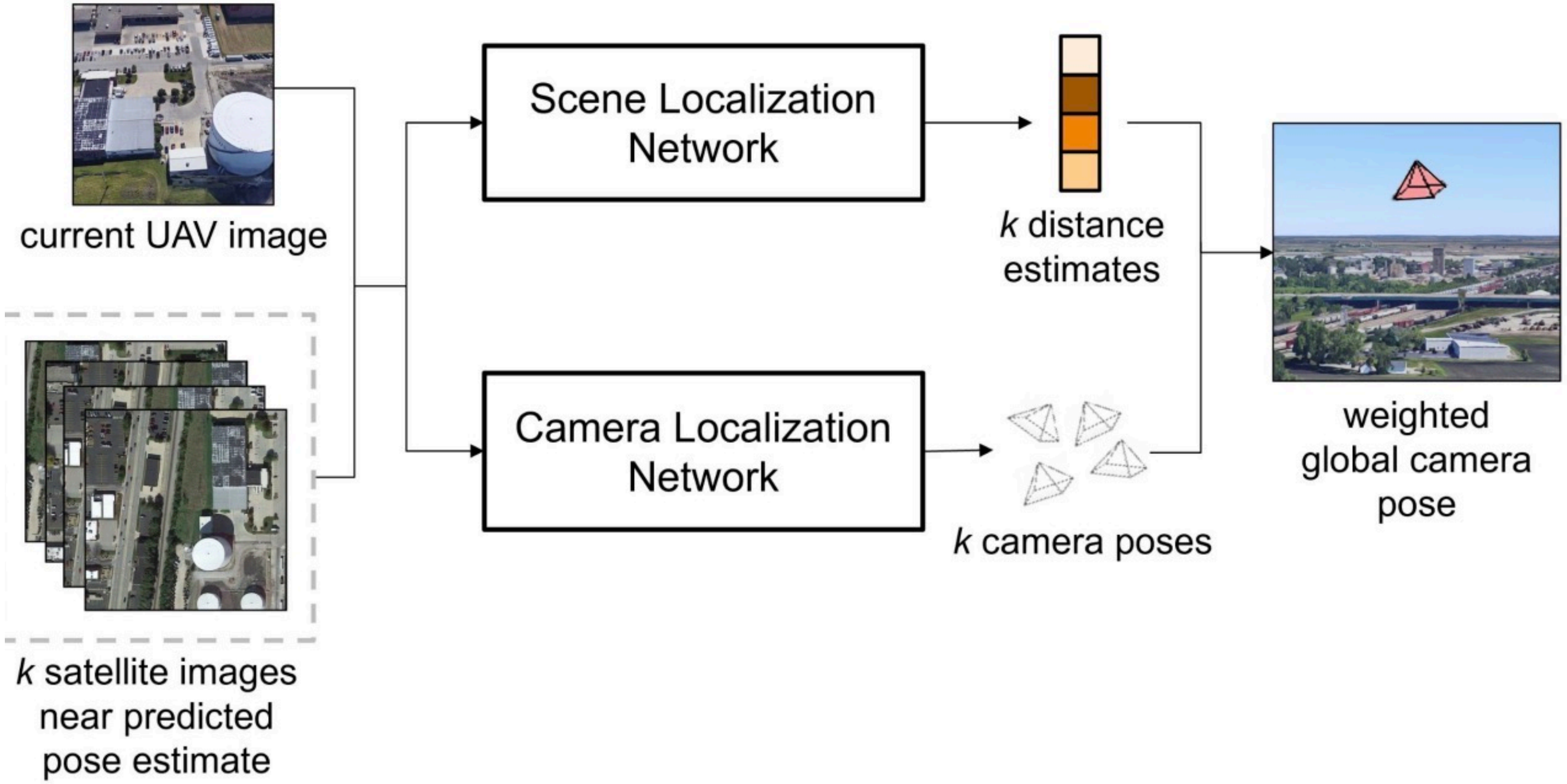
- Cross-View Policy Learning for Street Navigation



Li, Ang, et al. "Cross-view policy learning for street navigation." Proceedings of the IEEE International Conference on Computer Vision. 2019.

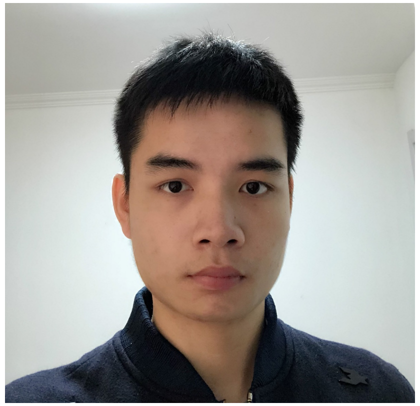
Why is image geo-localization important?

- UAV Pose Estimation using Cross-view Geo-localization

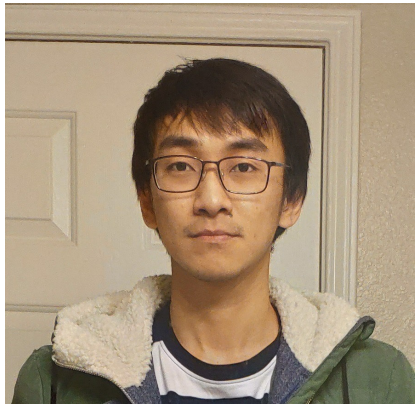


Shetty, Akshay, and Grace Xingxin Gao. "UAV Pose Estimation using Cross-view Geolocalization with Satellite Imagery." *arXiv preprint arXiv:1809.05979, ICRA 2019.*

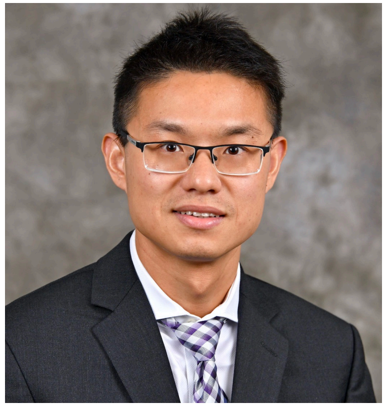
Street-to-Aerial View Matching for Image Geo-localization and Orientation Estimation



Sijie Zhu



Taojiannan Yang



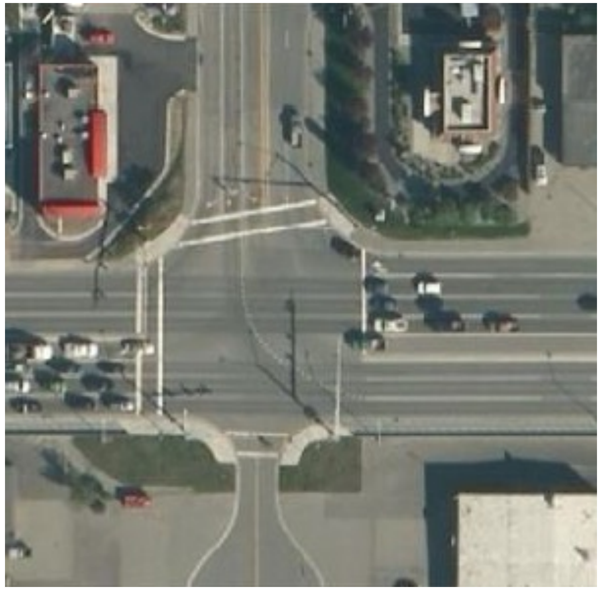
Chen Chen

[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", under review, IEEE Trans. on Image Processing, https://arxiv.org/pdf/1909.12977.pdf

Main Challenges

➢ Appearance Gap

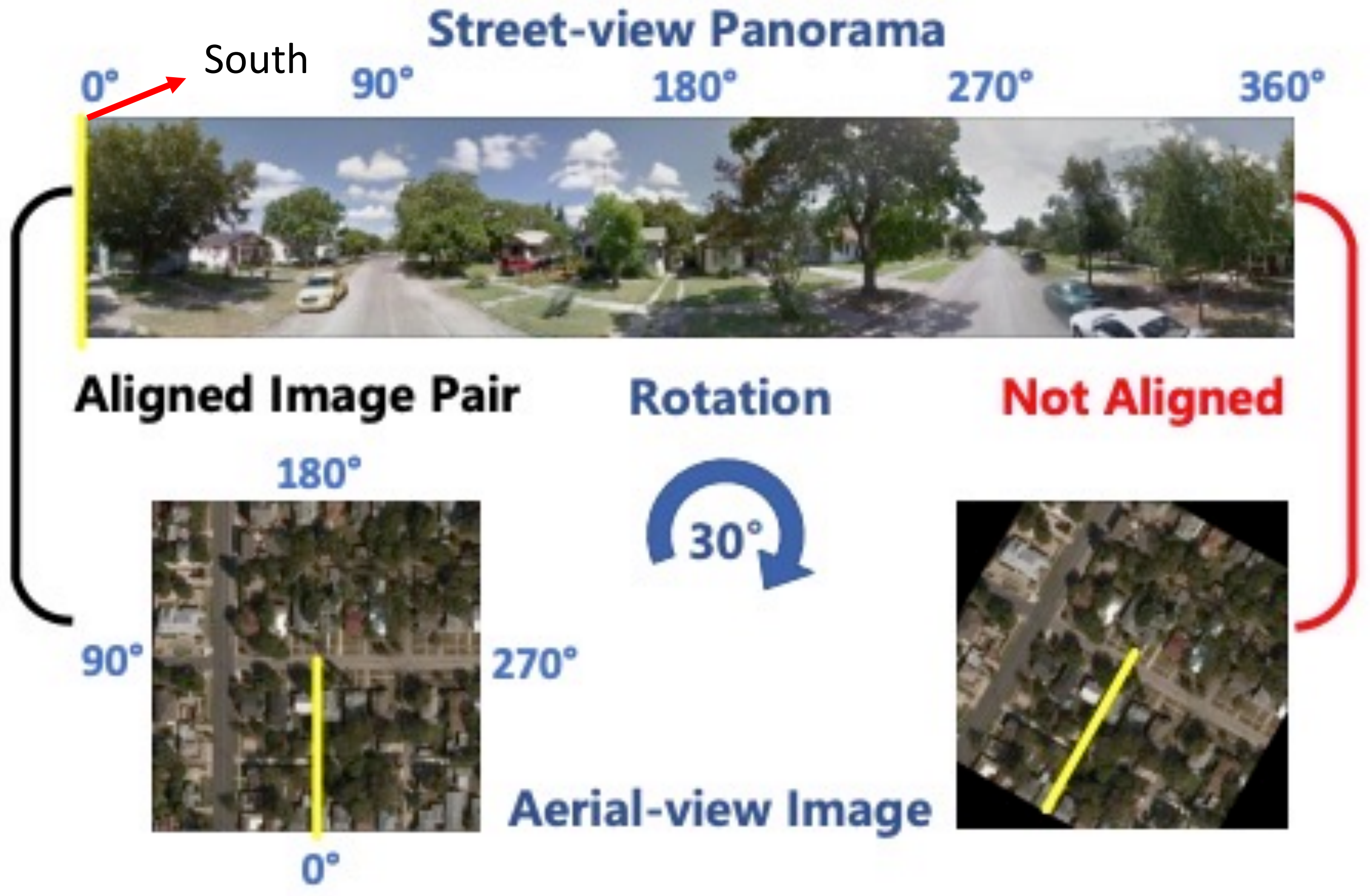


➢ Sample Imbalance

- The number of positive samples for an anchor street-view image is very limited in geo-localization, i.e., only one.

Main Challenges

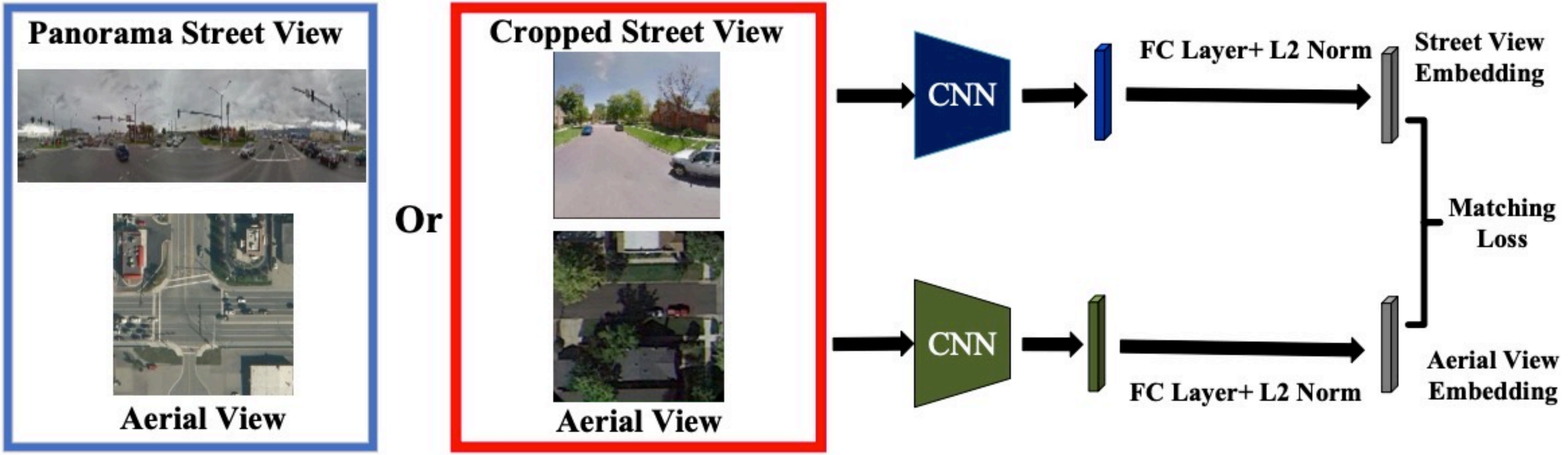
➢ Image alignment



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

Top-1 recall accuracy with different alignment settings

Overall Framework (network architecture)



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

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.

Geo-localization Examples

Geo-localization Examples

Geo-localization Examples

Street view id:18110

Aerial view rank:1
similarity:0.76

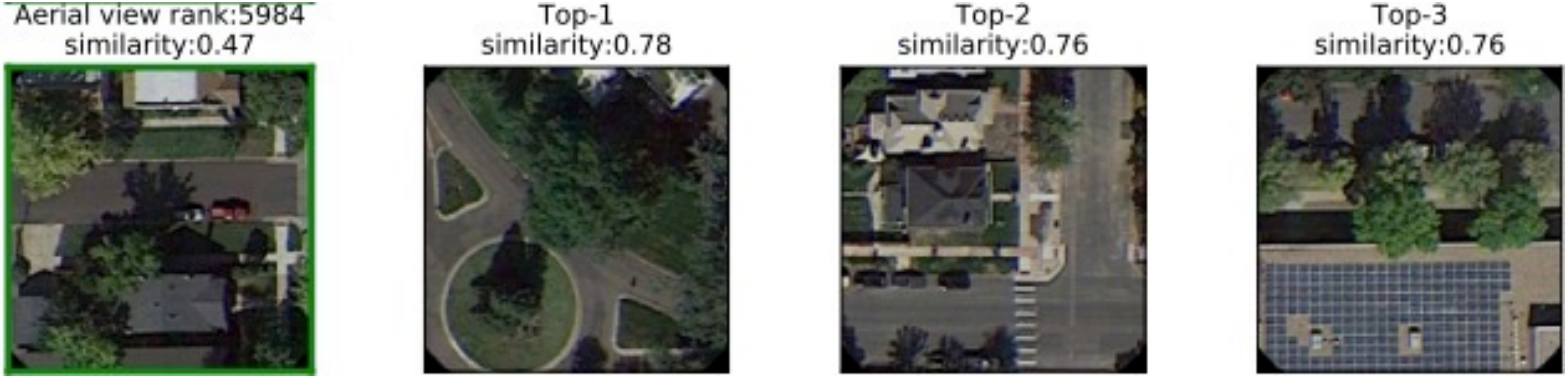
Top-1
similarity:0.76

Top-2
similarity:0.75

Top-3
similarity:0.74

Geo-localization Examples

A failure case



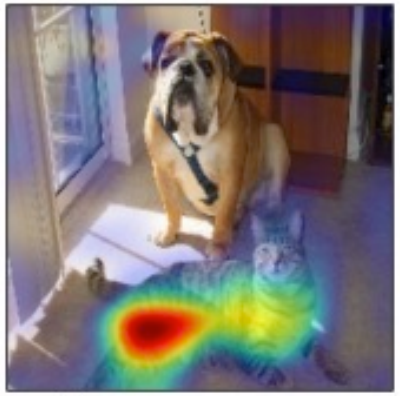
Visual Explanation of the Matching Results

- Visual explanation using Grad-CAM

What is Grad-CAM?

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

(a) Original Image



(c) Grad-CAM 'Cat'

(i) Grad-CAM 'Dog'

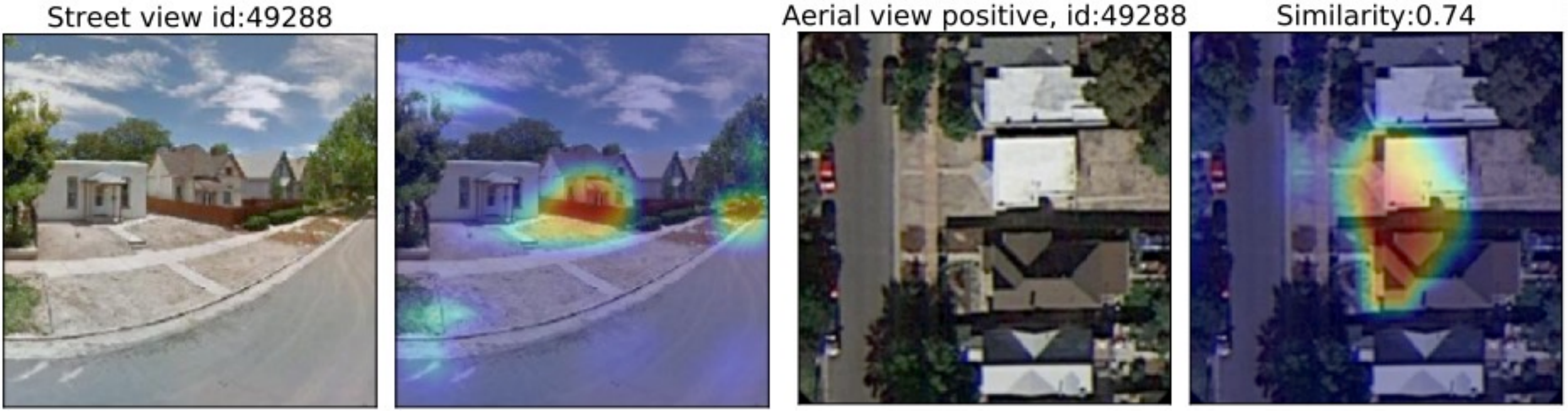
Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

Visual Explanation of the Matching Results

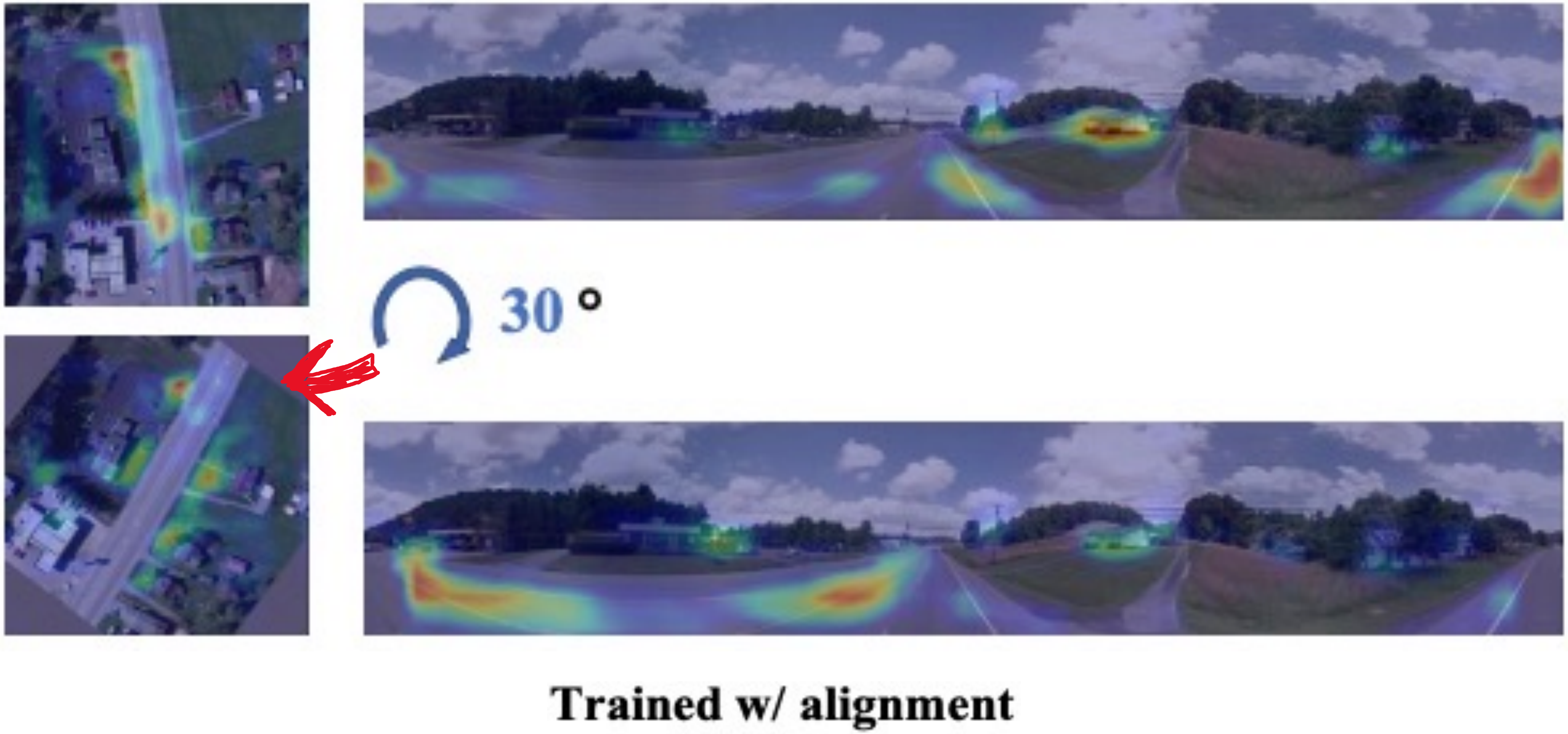
: regions that contribute the most to the similarity measure

The most activated regions are likely to be the same objects

Visual Explanation of the Matching Results

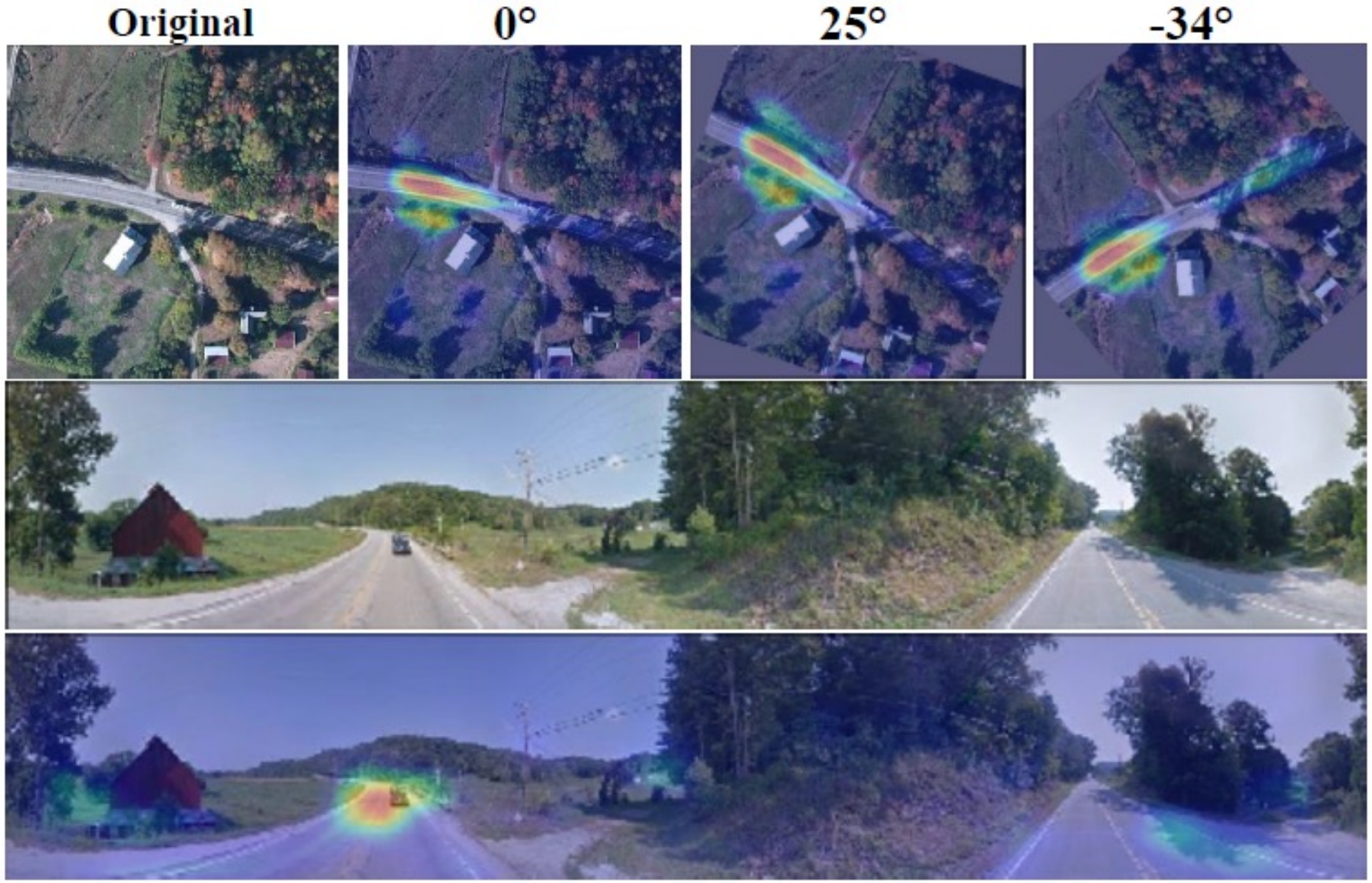


An Interesting Finding



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

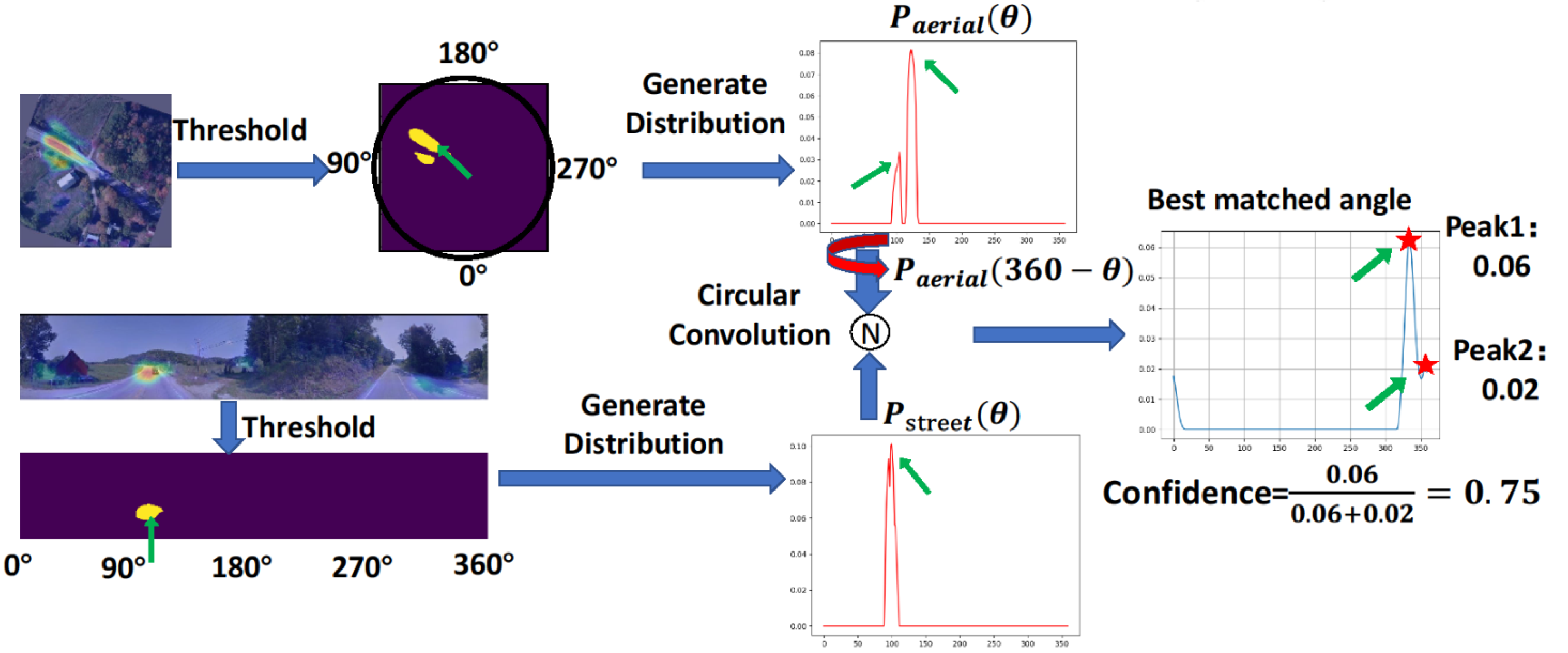
Orientation Estimation with Grad-CAM



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

Rotation Estimation Pipeline

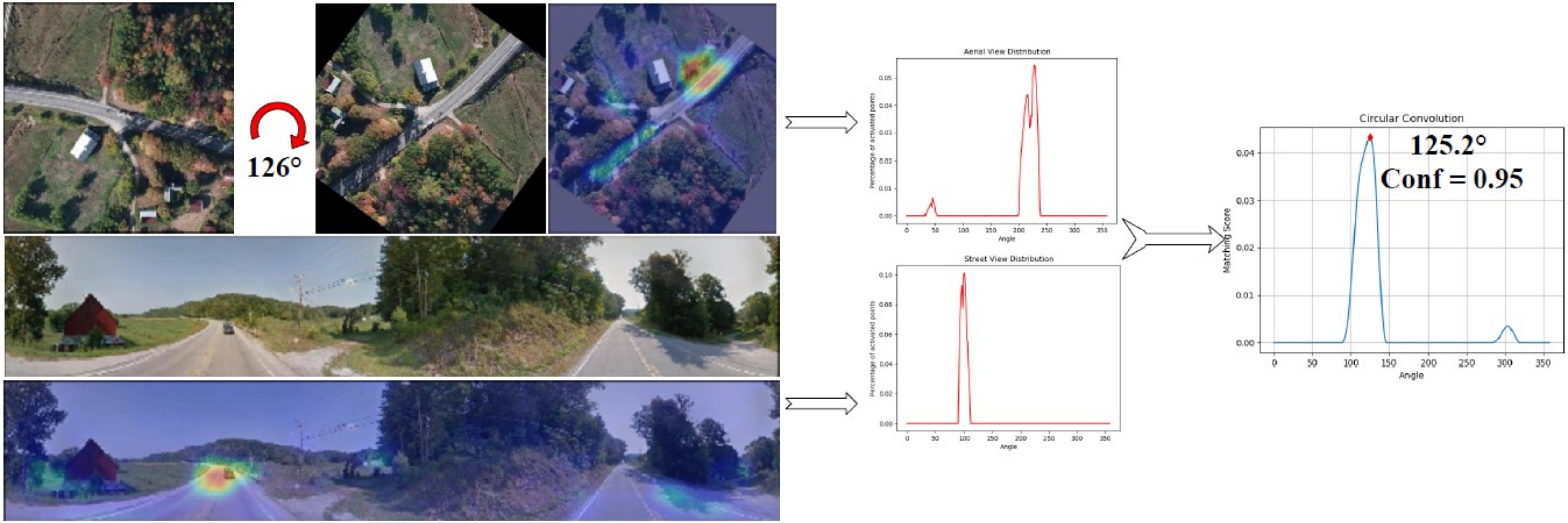
find the angle ϕ so that $p_{aerial}(\theta + \phi)$ best matches $p_{street}(\theta)$



$$\text{Confidence} = \frac{0.06}{0.06+0.02} = 0.75$$

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

Orientation Estimation Example



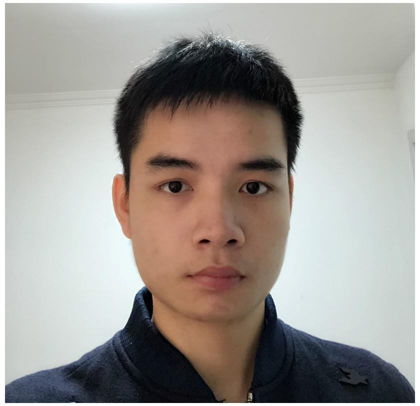
Summary

- Ablation study and visual explanation lead a key observation – the alignment has a great impact on the retrieval performance

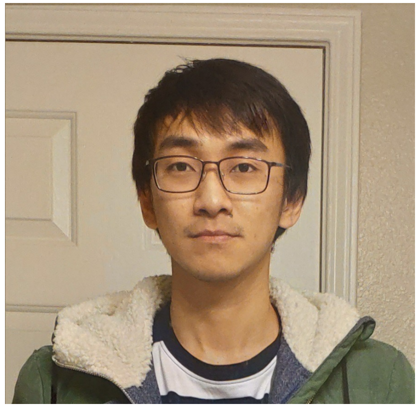
- We show that improvements on metric learning techniques boost the retrieval performance

- We discover that the orientation information between cross-view images can be estimated when the alignment is unknown

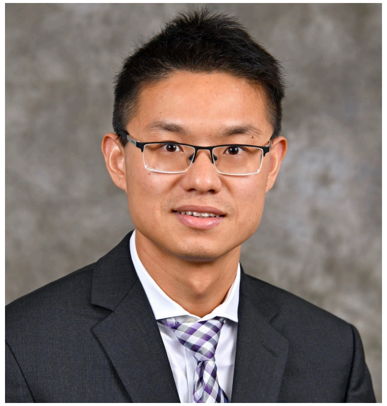
VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval



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

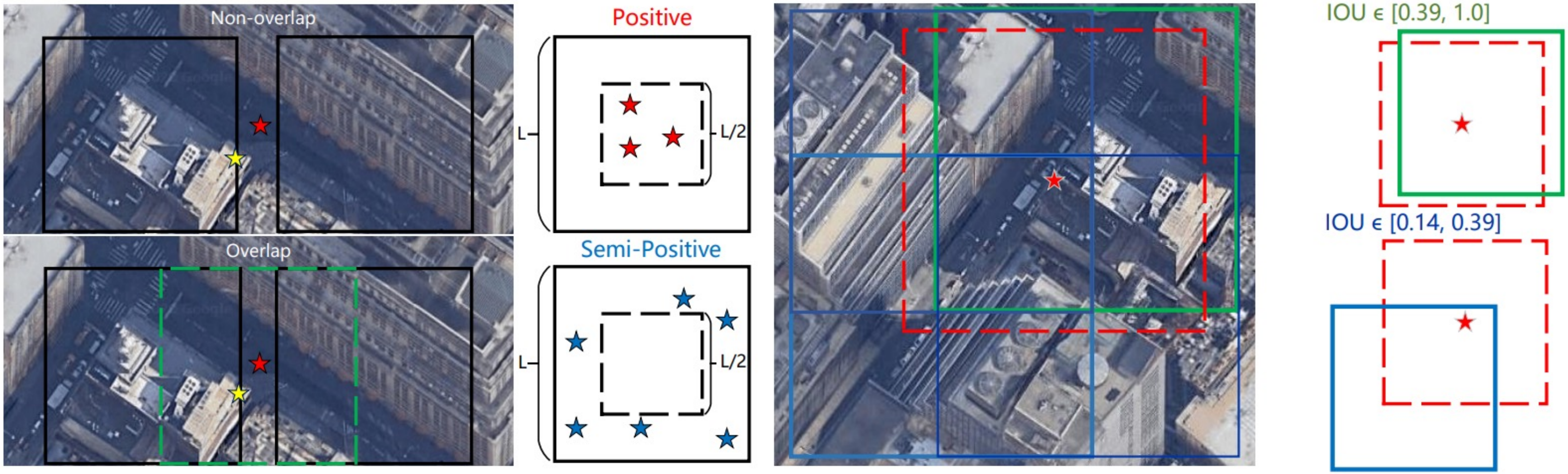
One-to-one Retrieval

- 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.

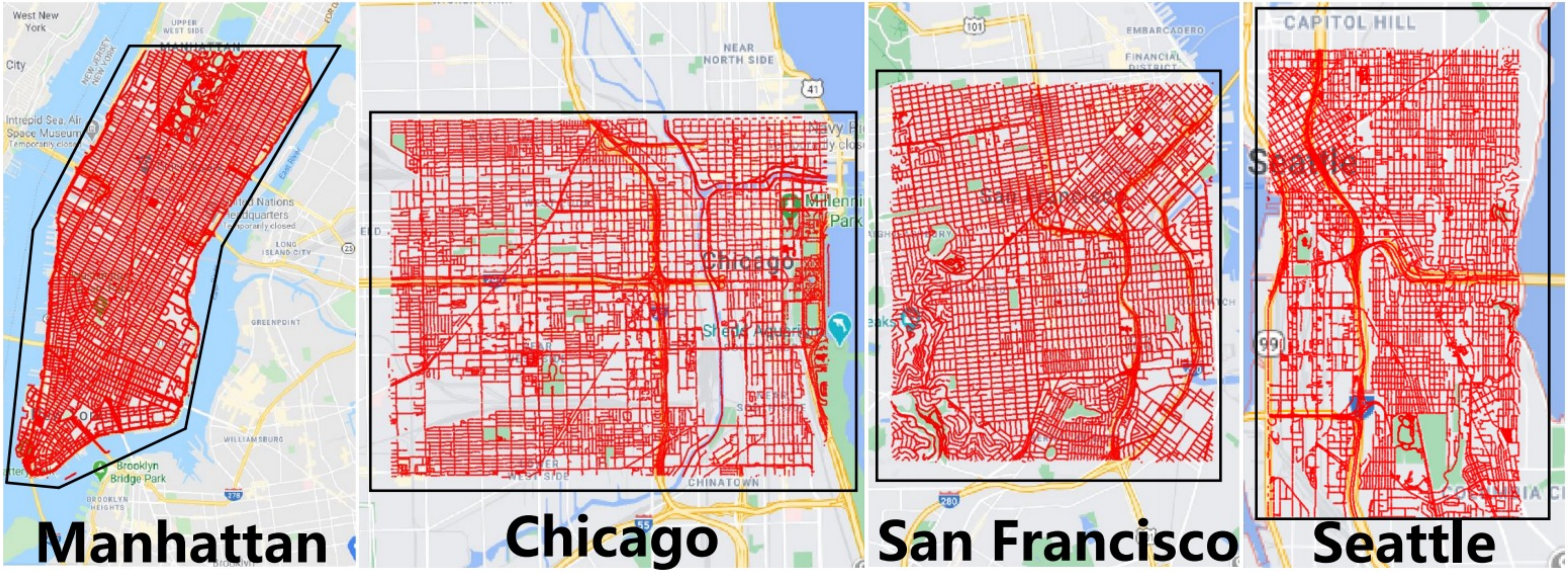
VIGOR

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) Positive vs Semi-positive. (c) Aligned (R) vs Positive (G) vs Semi-positive (B). (d) Positive vs Semi-positive IOU.

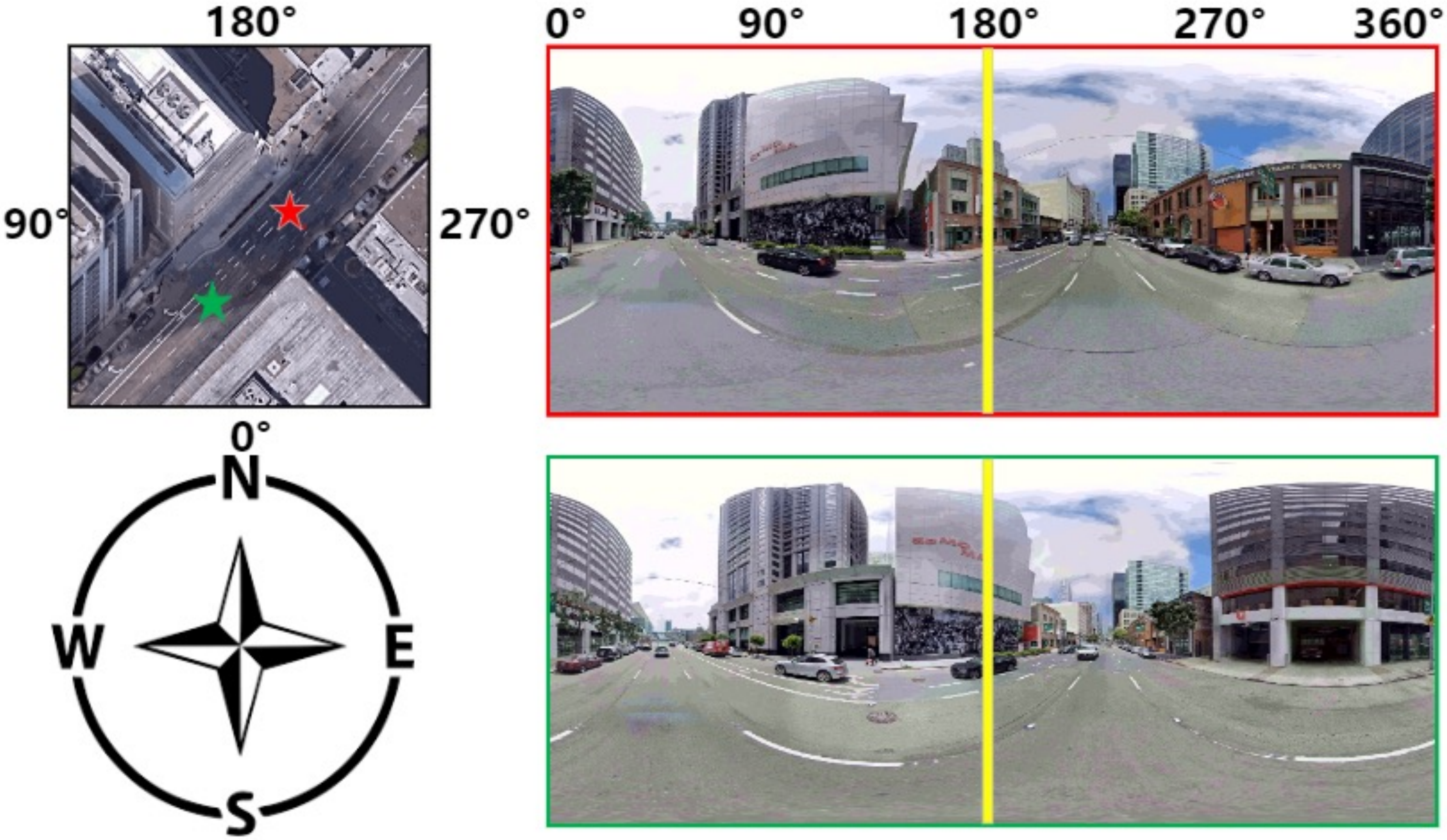
Data Distribution



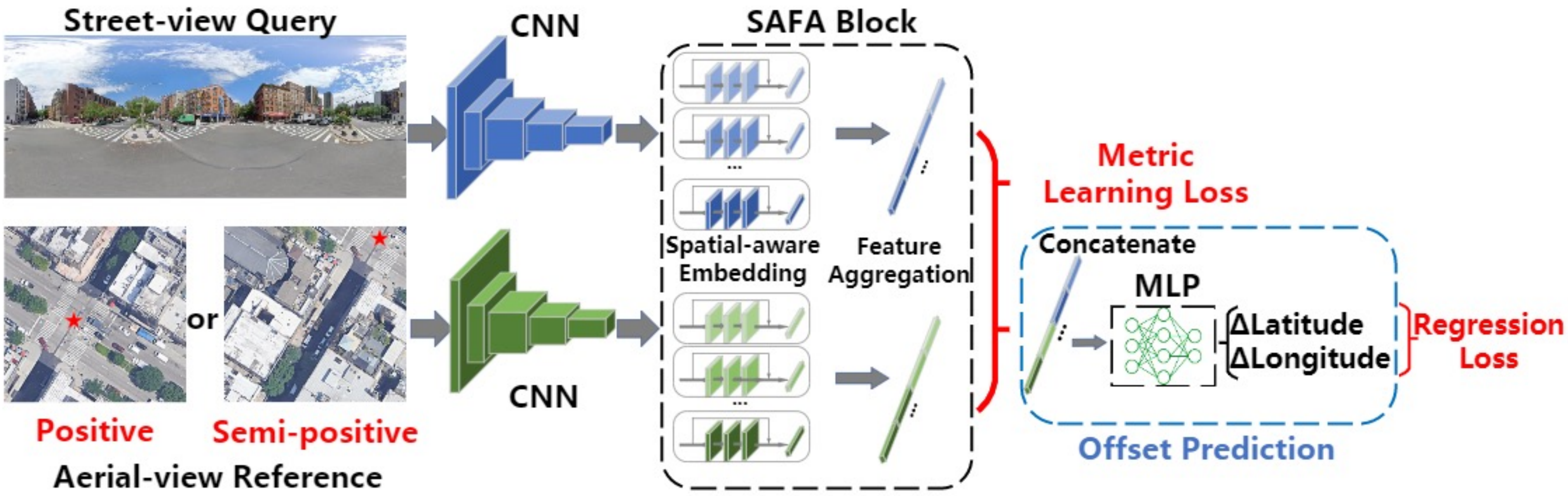
Datasets Comparison

	Vo [24]	CVACT [11]	CVUSA [27]	VIGOR (proposed)
Satellite images	$\sim 450,000$	$128,334$	$44,416$	$90,618$
Panoramas in total	$\sim 450,000$	$128,334$	$44,416$	$238,696$
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

Example Query and Reference



Coarse-to-fine Cross-view Localization



[SAFA] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial aware feature aggregation for image based cross-view geolocalization. In Advances in Neural Information Processing Systems, pages 10090–10100, 2019.

Beyond One-to-one

How to make use of the semi-positive images?

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

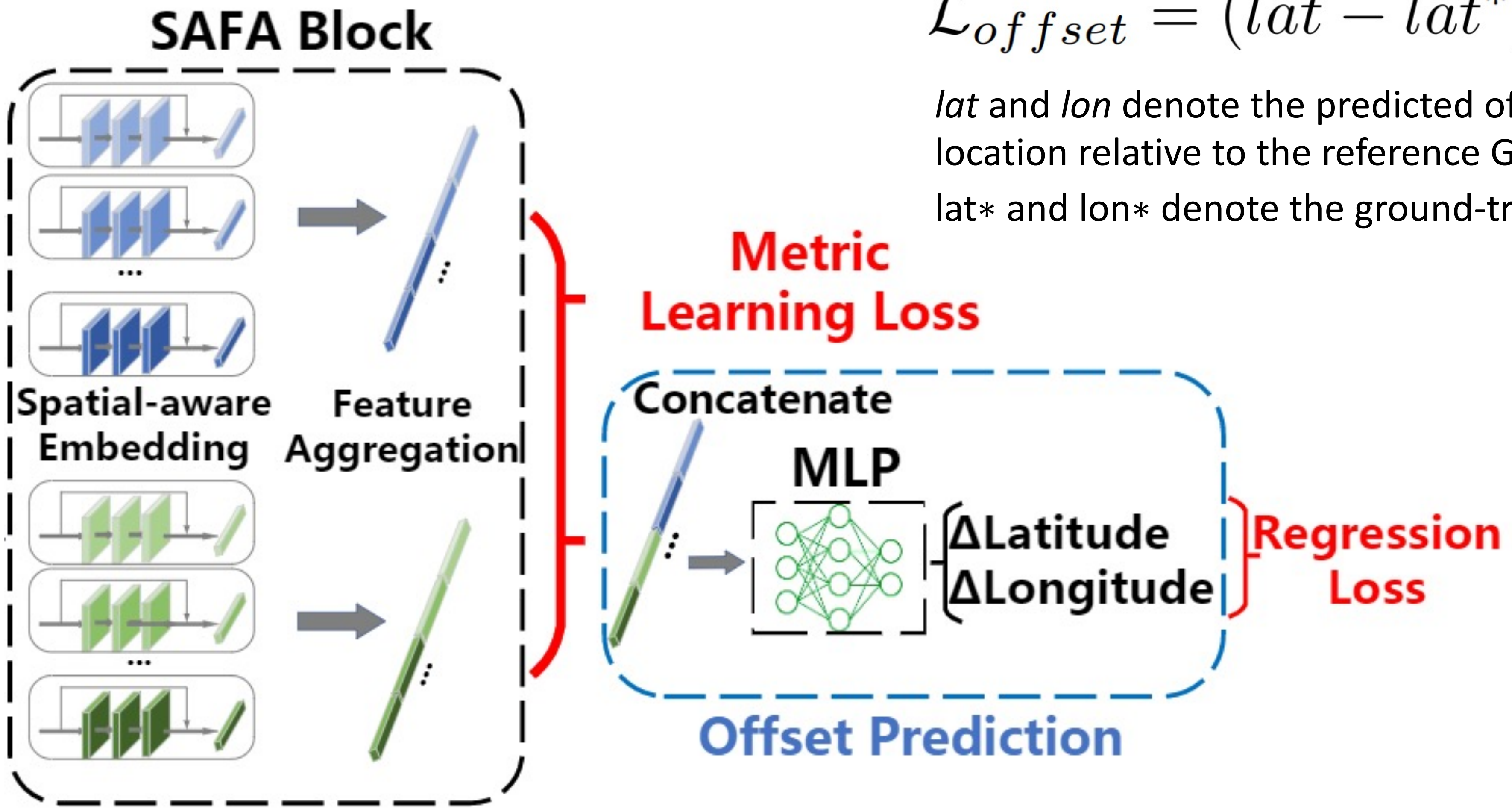
We force the ratio of the similarities in the embedding space to be close to the ratio of IOUs.

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.e.* 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



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

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

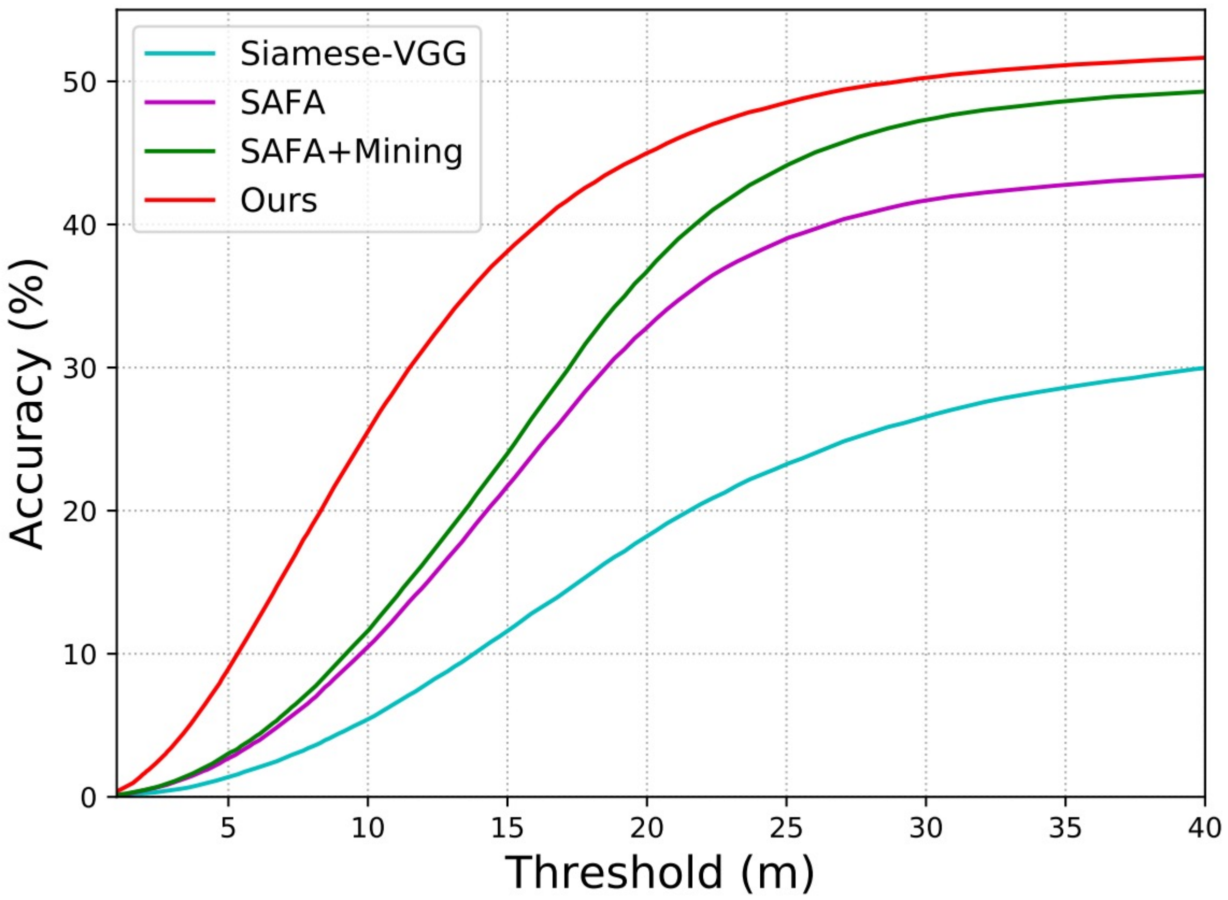
- Retrieval Performance

	Same-Area				Cross-Area			
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
Siamese-VGG ($\mathcal{L}_{triplet}$)	18.1	42.5	97.5	21.2	2.7	8.2	61.7	3.1
SAFA ($\mathcal{L}_{triplet}$)	33.9	58.4	98.2	36.9	8.2	19.6	77.6	8.9
SAFA+Mining (baseline, $\mathcal{L}_{triplet}$)	38.0	62.9	97.6	41.8	9.2	21.1	77.8	9.9
Ours ($\mathcal{L}_{hybrid}$)	**41.1**	**65.8**	**98.4**	**44.7**	**11.0**	**23.6**	**80.2**	**11.6**

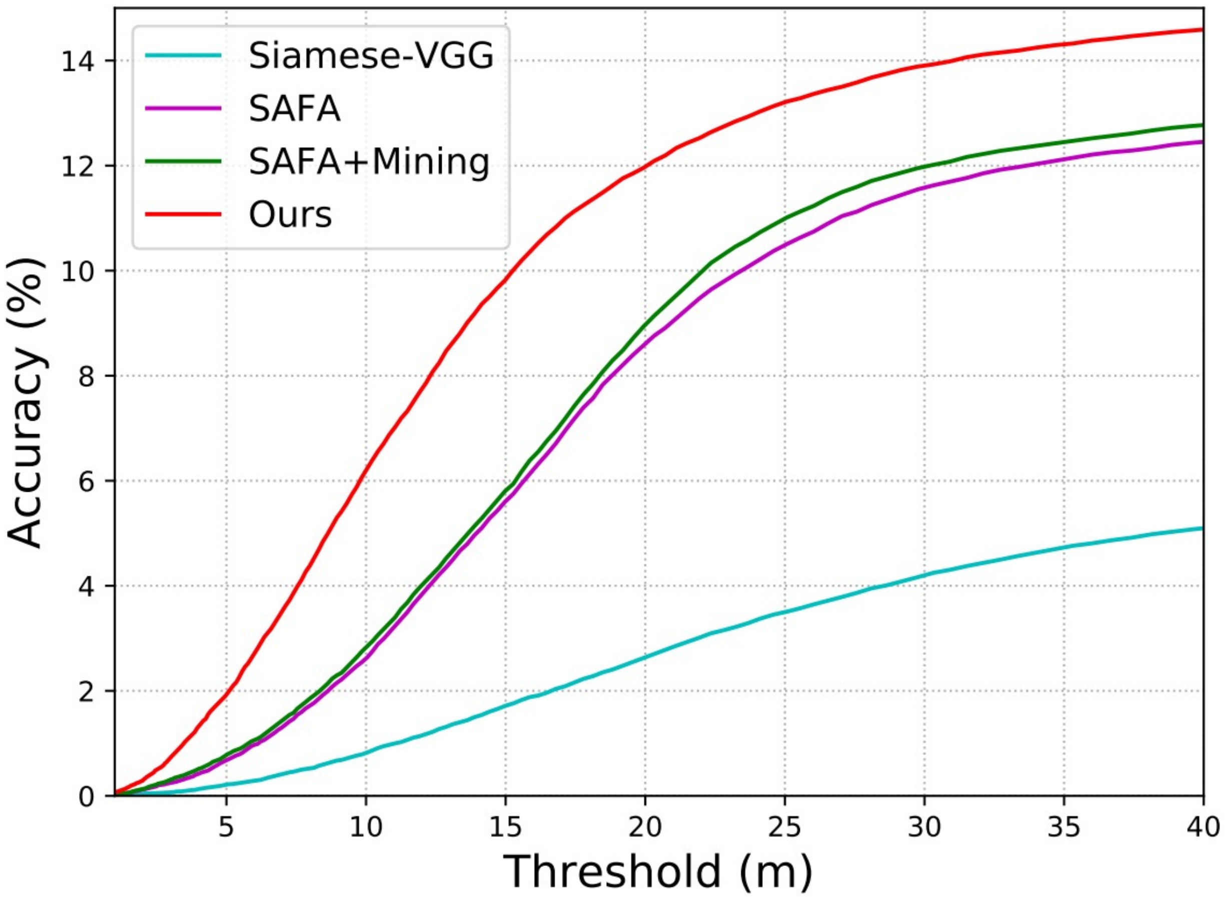
Comparison with State-of-the-art

- Localization in terms of meters

Same-area

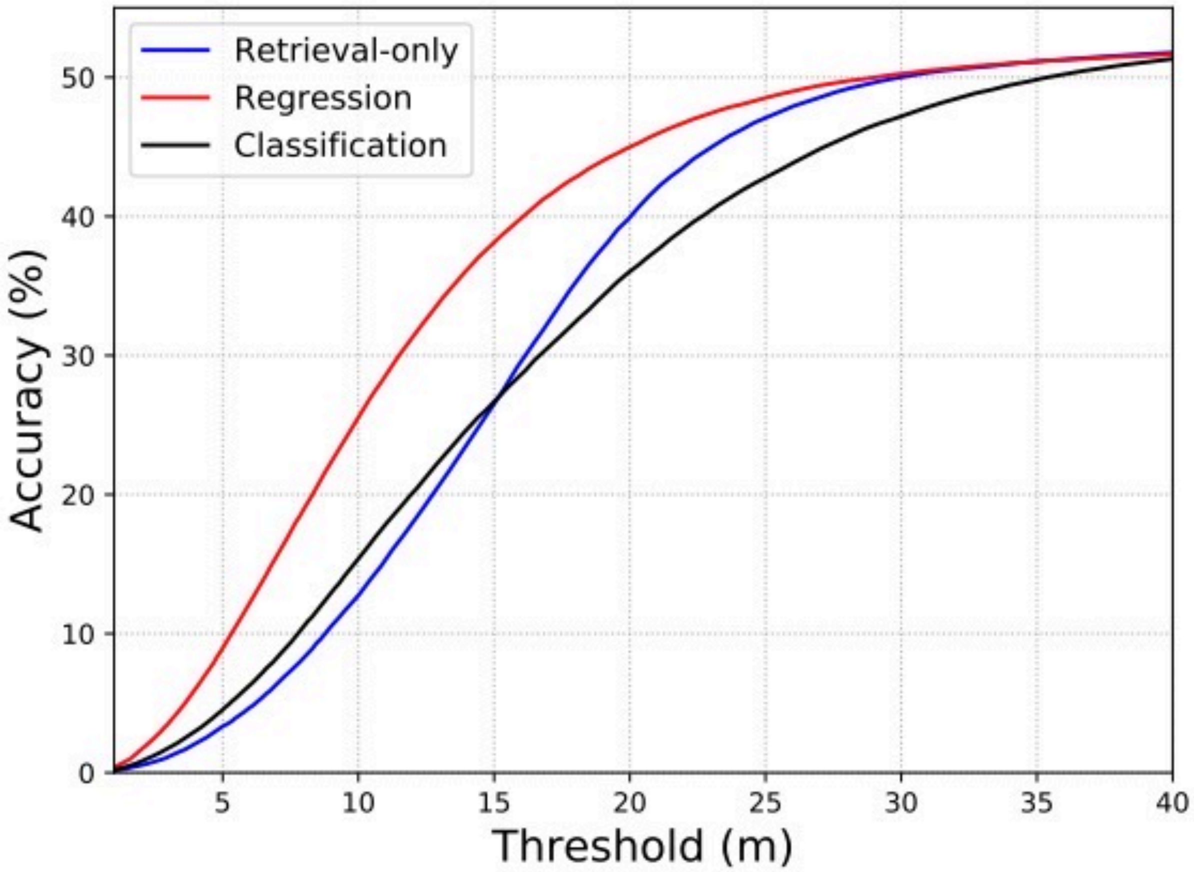


Cross-area



The Effect of Offset Prediction

- Localization in terms of meters



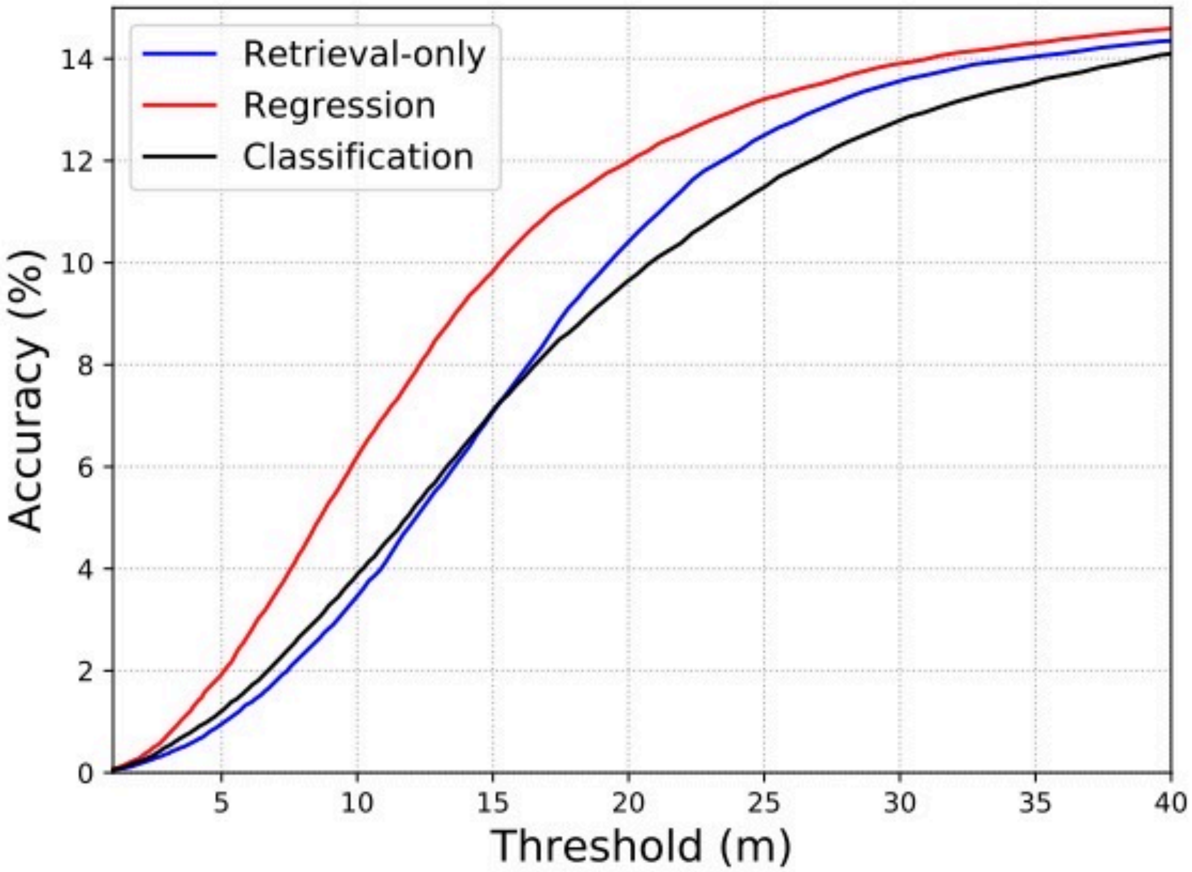


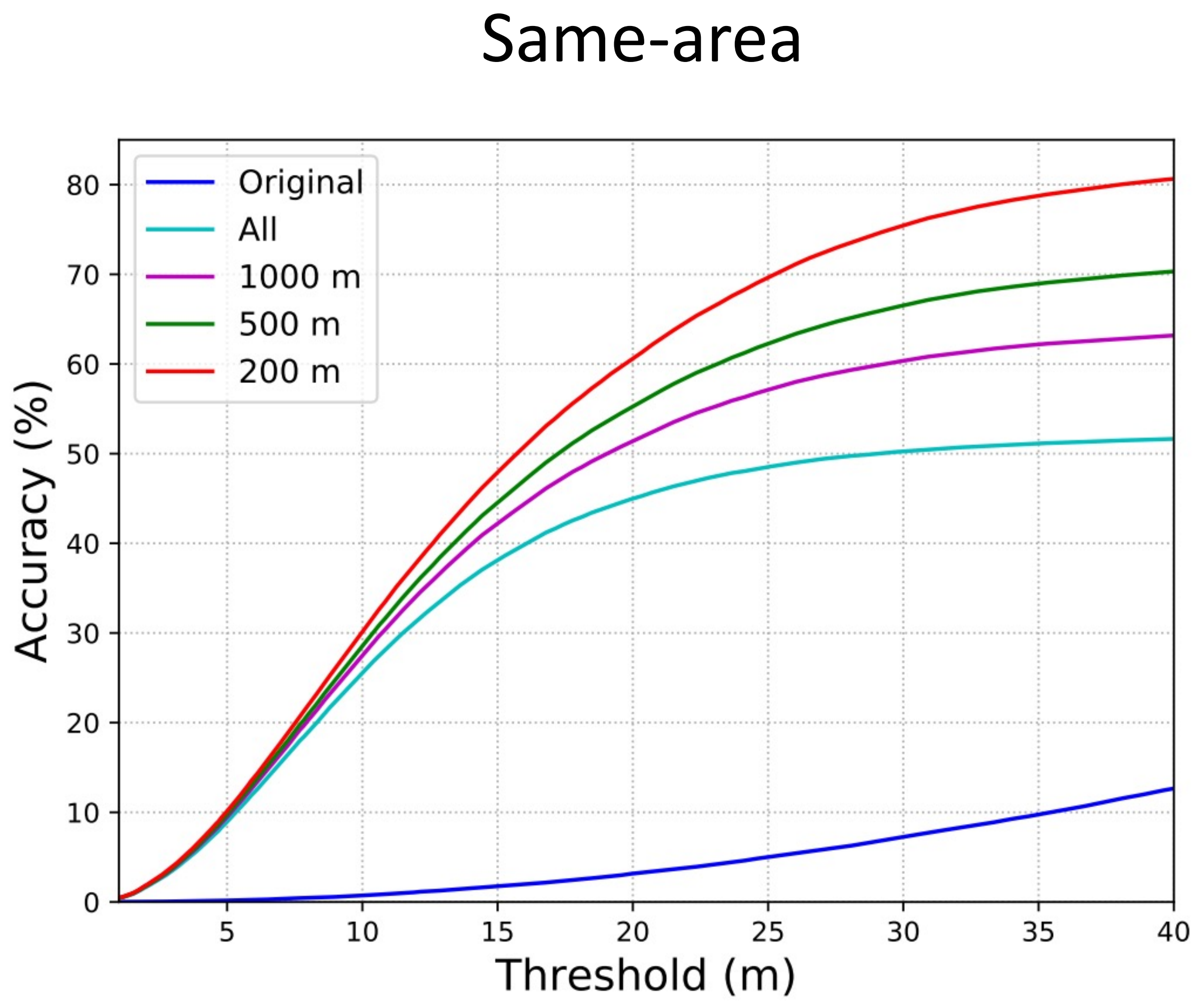
Figure 6. Same-area (left) and cross-area (right) meter-level localization accuracy of different offset prediction methods.

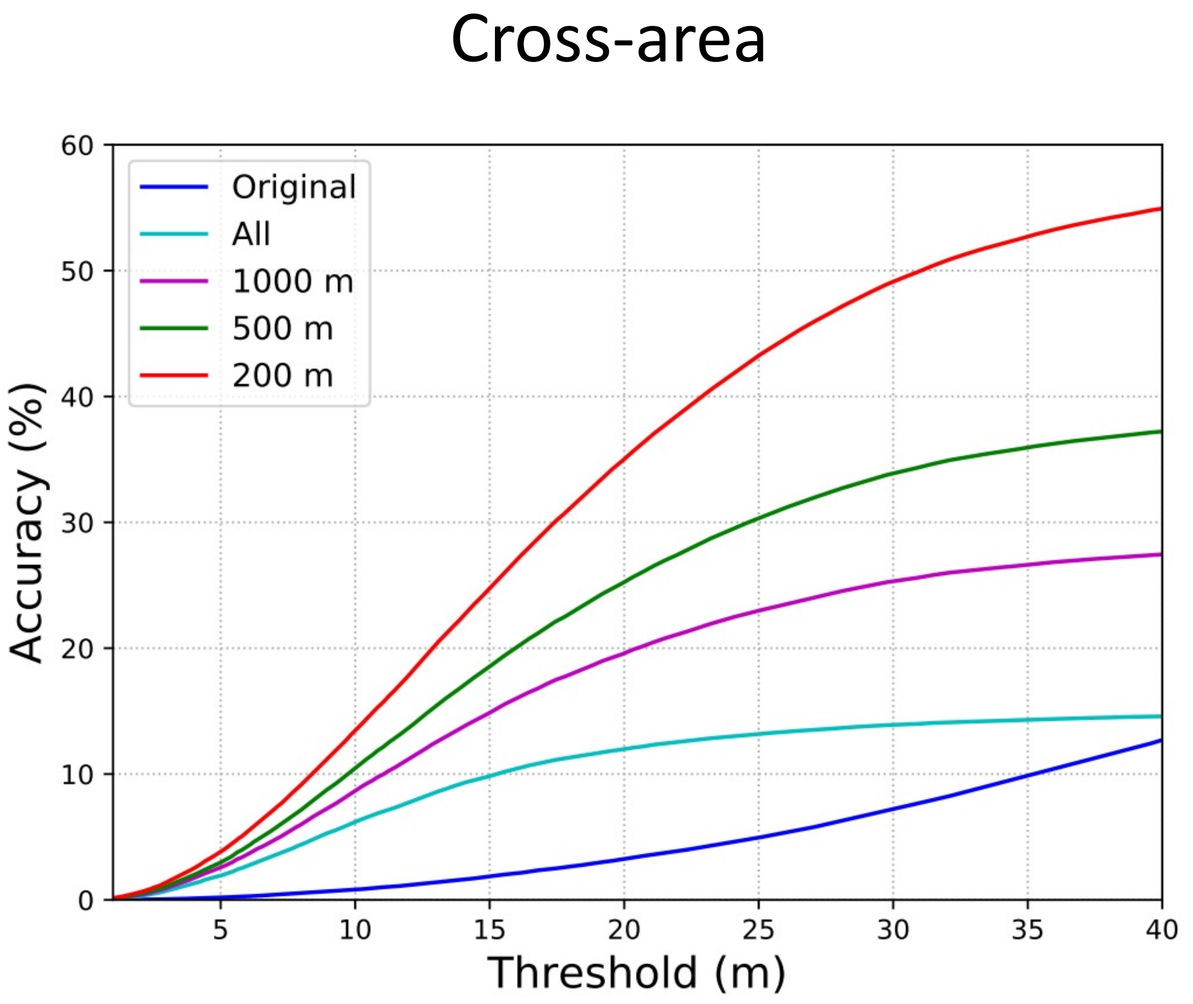
Noisy GPS Refinement

- 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 m	49.2	76.7	19.9	41.5
500 m	54.1	82.6	26.4	53.3
200 m	60.9	90.6	37.7	72.0

Noisy GPS Refinement





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

- We validate the potential of the proposed framework for noisy GPS refinement.

Website:

https://github.com/Jeff-Zilence/VIGOR

Thank you