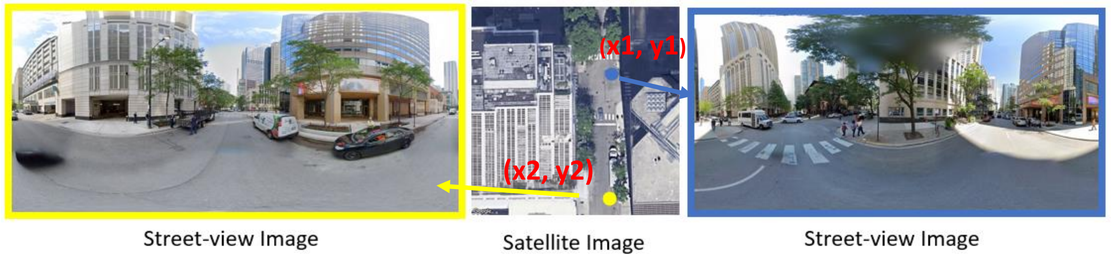
Cross-Attention Between Satellite and Ground Views for Enhanced Fine-Grained Robot Geo-Localization

Dong Yuan[1], Frederic Marine[1], Feras Dayoub[2,3]
[1]QUT Centre for Robotics, [2]Australian Institute for Machine Learning, [3]University of Adelaide

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Fine-Grained Cross-View Geo-Localization

- **Cross-view image pairs:** a satellite image covers more than one street-view image, and the street-view images are located far away from the center of the satellite image.
- **Fine-grained location prediction:** given a street-view image, predict satellite image pixel coordinates corresponding to this street-view image.



Our Contributions

1. Leverage the cross-attention mechanism to compute the cross-correlation between aerial and street views.
2. Combine classification prediction and regression prediction to address the cross-view fine-grained localization problem.
3. Outperforms the SOTA method (MCC) [2] in fine-grained location prediction on the VIGOR benchmark [3].

Cross-Attention Between Ground-Satellite Views

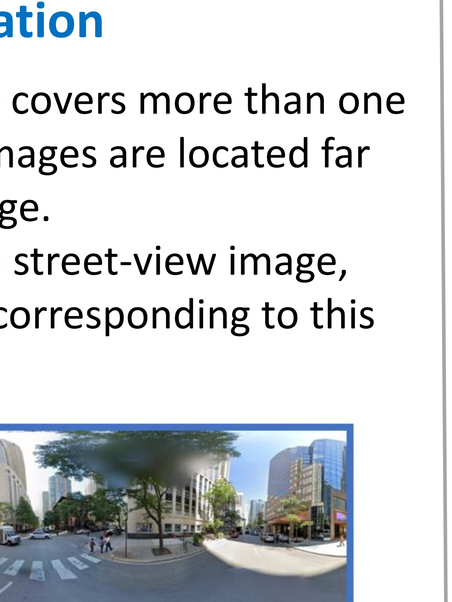
- Features f_s, and f_g, are from satellite and ground views separately.
- Feature f_s, is projected to Key (K_s) and Value (V_s), and feature f_g, is projected to Query (Q_g).
- Compute cross-attention score:

$$= \mathrm{Softmax}(Q_g^T \times K_s)$$

- Obtain the new ground-view feature:

$$f_g', = \mathrm{AttentionScore} \times V_s$$

- Feature f_g', fuses relevant information from both the satellite and ground views.



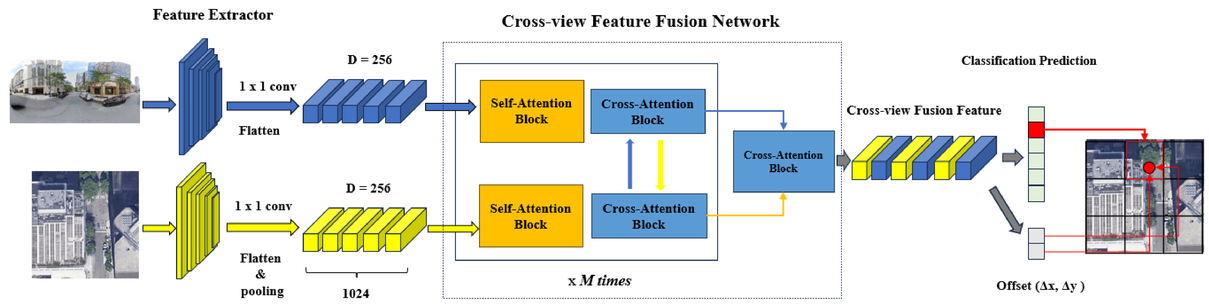
Cross-Attention Block

Location Prediction

- A classification header predicts the current street-view image is in which cell of the $N \times N$ grid.
- A regression header predicts the pixel coordinate offset based on classification results.

Proposed Architecture

- Three components: local feature extractors, cross-view feature fusion network and prediction headers.



Experiment Results

- Results on VIGOR Dataset

Model	Same-Area				Cross-Area			
	Positives		Pos+Semi-Pos		Positives		Pos+Semi-Pos	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
VIGOR	10.55	9.31	16.64	13.82	11.26	10.02	18.66	16.73
MCC	9.86	4.58	13.45	5.39	13.06	6.31	17.13	7.78
Ours	**6.72**	**2.68**	**9.32**	**3.11**	**7.28**	**2.83**	**9.78**	**3.89**

- Results on Oxford RobotCar Dataset [1]

Model	Mean	Median
VIGOR	2.29±0.31	1.72±0.21
MCC	1.77 ±0.25	**1.24±0.10**
Ours	**1.77±0.20**	1.32±0.08

Model Error	Test 1	Test 2	Test 3	Average
MCC mean	**1.42**	**1.95**	1.94	1.77 ± 0.25
Ours mean	1.50	1.97	**1.83**	**1.77±0.20**
MCC median	**1.10**	**1.33**	**1.29**	**1.24 ± 0.10**
Ours median	1.22	1.42	1.33	1.32 ± 0.08

Conclusions

- The cross-attention mechanism can establish better relations between ground and aerial views.
- The combination of classification and regression prediction achieves more accurate fine-grained localization.
- Our approach reduces the median localization distance error by 43% and 50% respectively in the same area and unseen areas on the VIGOR benchmark

1. Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. The International Journal of Robotics Research (IJRR), 36(1):3– 15, 2017.
2. Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In ECCV, pages 90–106. Springer, 2022.
3. Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In CVPR, pages 3640–3649, 2021.