Cross-View Geo-Localization: Ground-to-Aerial Image Matching

Mubarak Shah
shah@crcv.ucf.edu
Center for Research in Computer Vision
University of Central Florida
http://www.crcv.ucf.edu

Geo-Localization

- Pixel-Wise Geo-Localization
  - Given a query image, geo-localize each pixel by aligning an image with the geodetically accurate reference image.

- Image-Based Geo-Localization
  - Given a query image, determine its GPS location by matching it with geo-tagged reference images.

Contents

- Pixel-Wise Geo-localization
  - Geodetic Alignment of Aerial Video Frames
- Image-Based Geo-Localization
  - Same View (Street-View to Street-View)
    - Generalized Maximum Clique (PAMI, 2014)
    - Constraint Dominant Sets (PAMI, 2017)
  - Cross-View Geo-Localization
    - Bird’s Eye-View to Street View (CVPR, 2017)
    - Aerial to Ground View (ICCV, 2019)

Geodetic Alignment of Aerial Video Frames

Y. Sheikh, S. Khan, M. Shah and R. Cannata

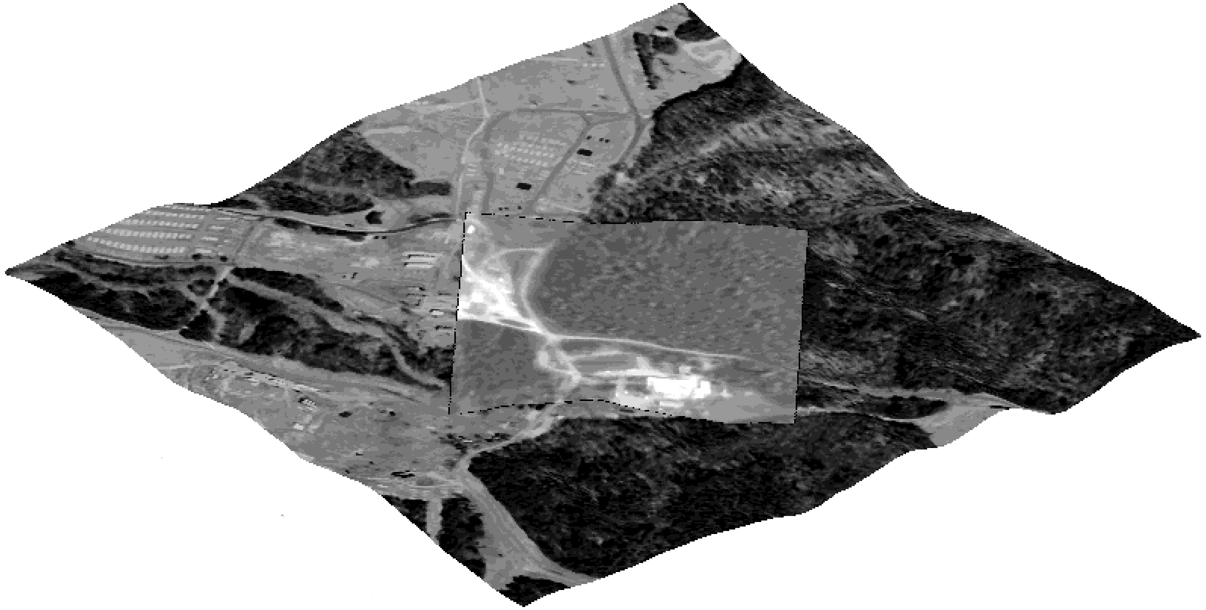
2003

Associate Professor
CMU/FaceBook

geoRegistration

Data Overview

- Aerial Video Data
- Reference Data
- Telemetry



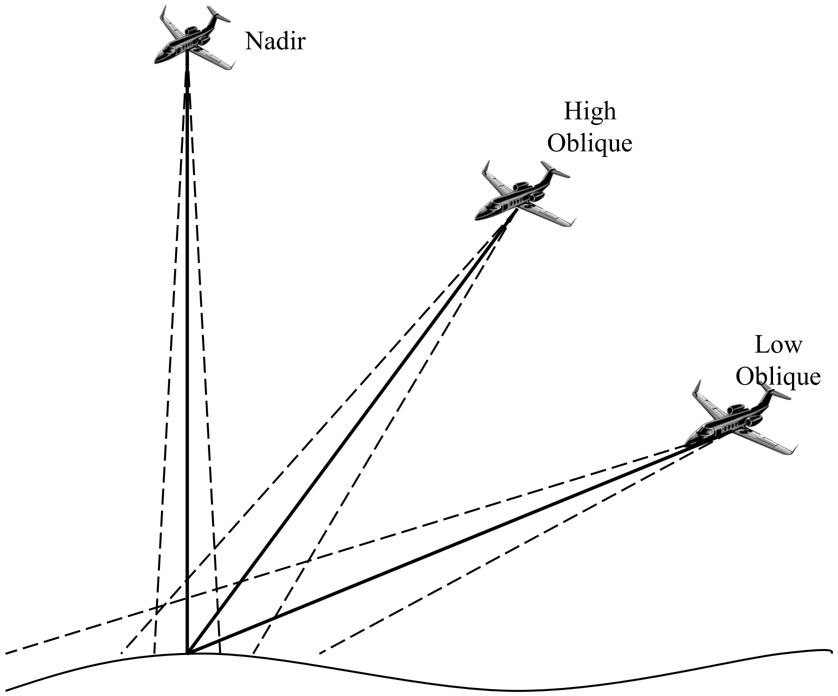
*geo*Registration

Different Viewpoints

Nadir

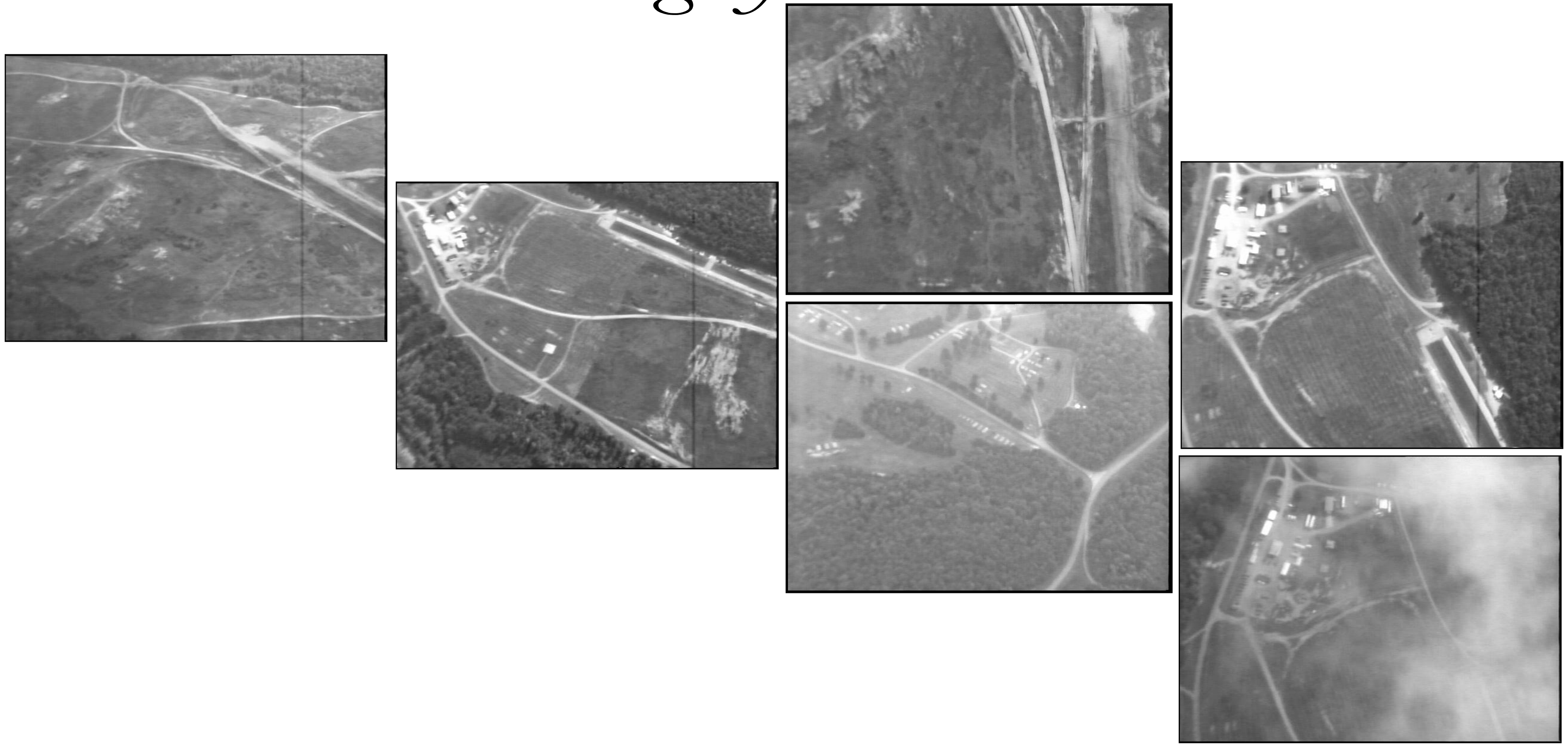
High Oblique

Low Oblique



geoRegistration

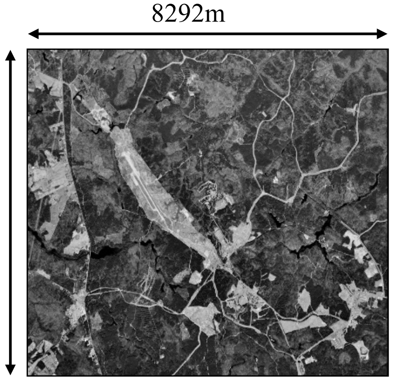
Aerial Video Imagery



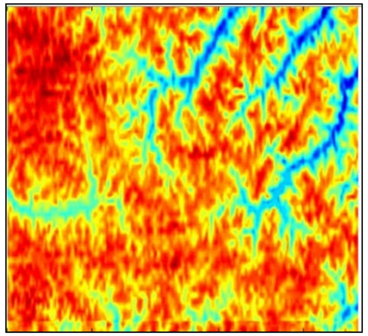
*geo*Registration

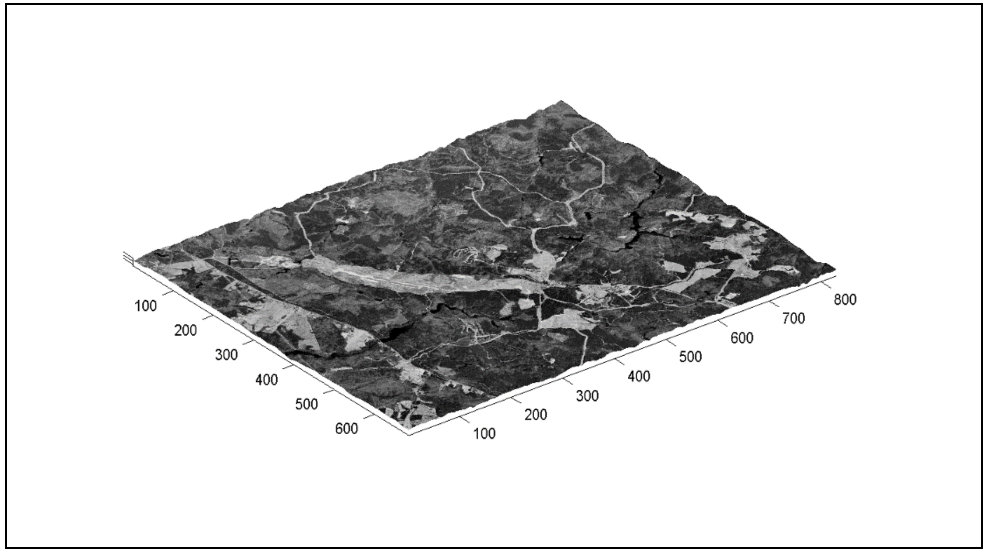
Reference Data

DOQ
(Digital Ortho Quad



DEM
(Digital Elevation Map)

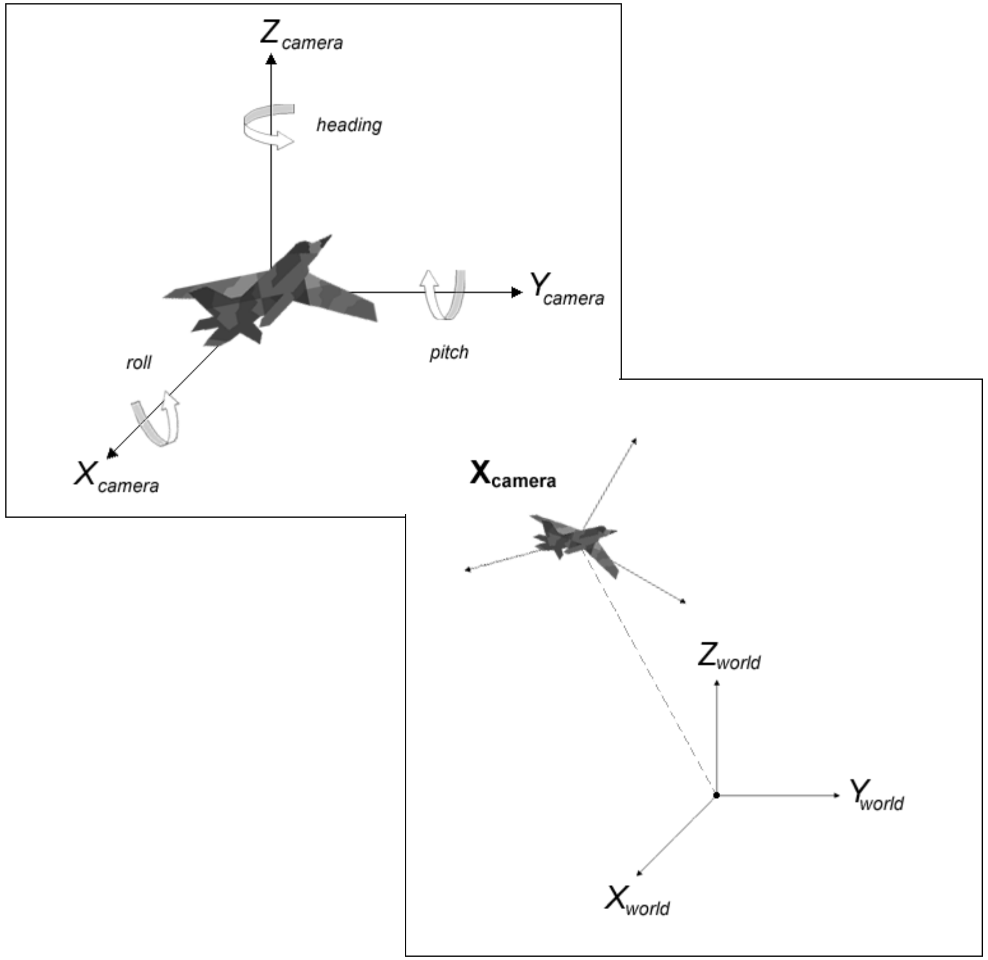




*geo*Registration

Telemetry Data

- Vehicle Longitude
- Vehicle Latitude
- Vehicle Height
- Vehicle Heading
- Vehicle Roll
- Vehicle Pitch
- Camera Elevation Angle
- Camera Scan Angle
- Camera Focal Length

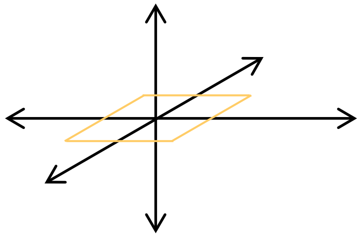


Sensor Model

$$\mathbf{X}_{\mathbf{camera}} = [X_{camera} \quad Y_{camera} \quad Z_{camera}]^{\mathrm{T}}$$

$$\mathbf{X}_{\mathbf{world}} = [X_{world} \quad Y_{world} \quad Z_{world}]^{\mathrm{T}}$$

$$\mathbf{X}_{\mathbf{camera}} = \Pi_t \mathbf{X}_{\mathbf{world}}$$



$\Pi_t = f$ (camera elevation, camera scan, vehicle pitch, vehicle roll, vehicle heading, vehicle elevation)

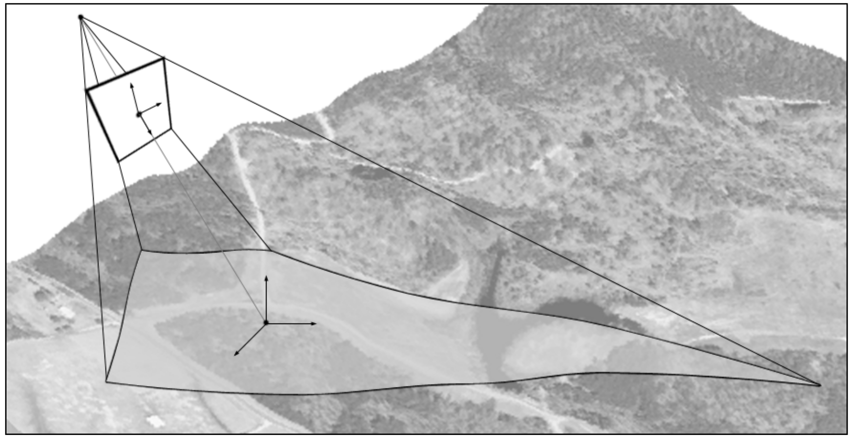
$$\Pi_t = \begin{bmatrix} \cos\omega & 0 & -\sin\omega & 0 \\ 0 & 1 & 0 & 0 \\ \sin\omega & 0 & \cos\omega & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\tau & \sin\tau & 0 & 0 \\ -\sin\tau & \cos\tau & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\phi & 0 & -\sin\phi & 0 \\ 0 & 1 & 0 & 0 \\ \sin\phi & 0 & \cos\phi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\beta & \sin\beta & 0 \\ 0 & -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\alpha & \sin\alpha & 0 & 0 \\ -\sin\alpha & \cos\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \Delta T_x \\ 0 & 1 & 0 & \Delta T_y \\ 0 & 0 & 1 & \Delta T_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Sensor Model

$$\Pi_c = P\Pi_t$$

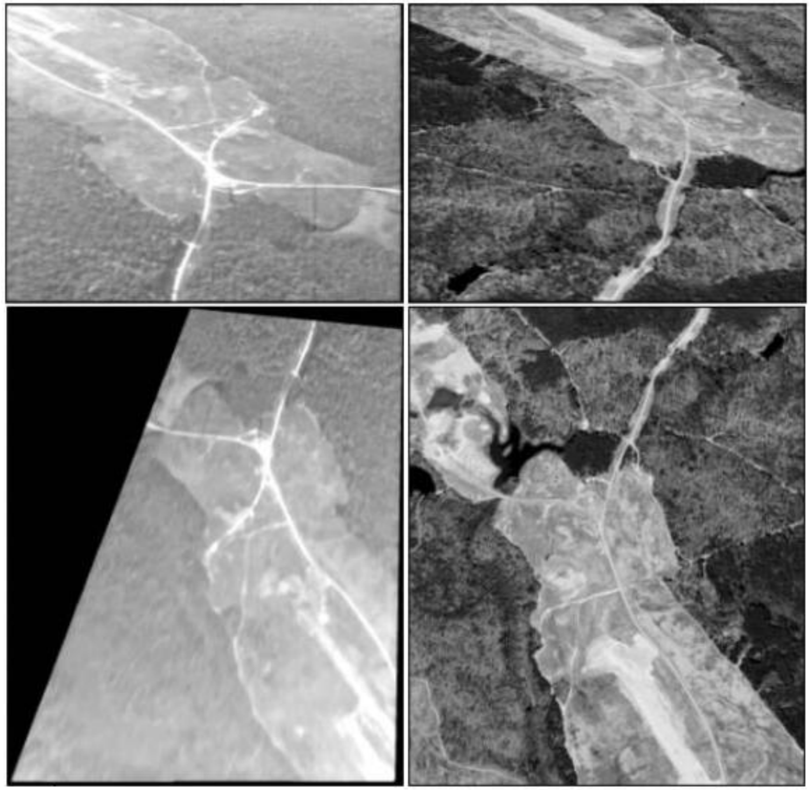
$$\Pi_c = PG_yG_zR_yR_xR_zT$$

$$\mathbf{x}_{\mathbf{video}} = \Pi_c\mathbf{X}_{\mathbf{world}}$$



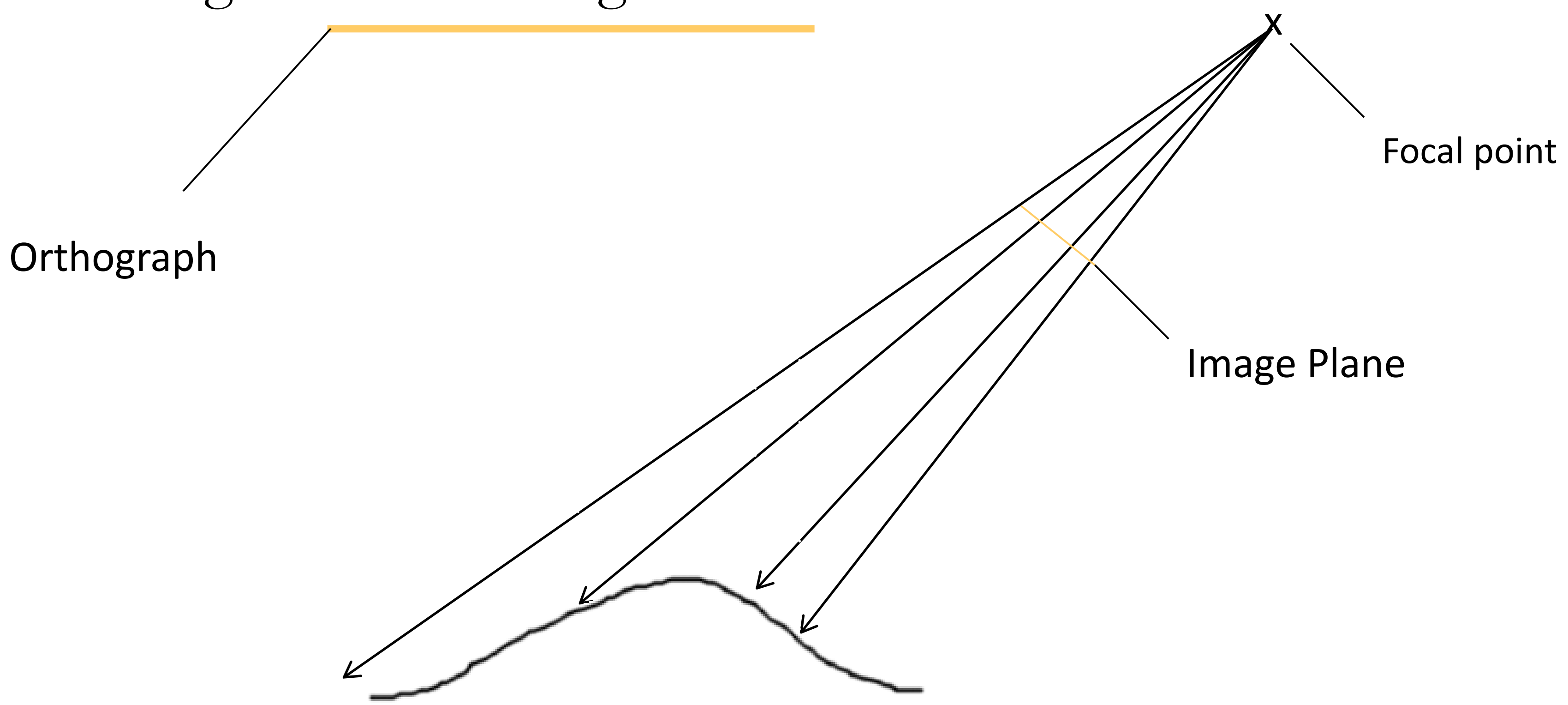
*ortho*Rectification

- Bringing both imageries onto a common view projection (**Bridging the gap**)
- Accounts for gross misalignment
- Not accurate due to telemetry noise



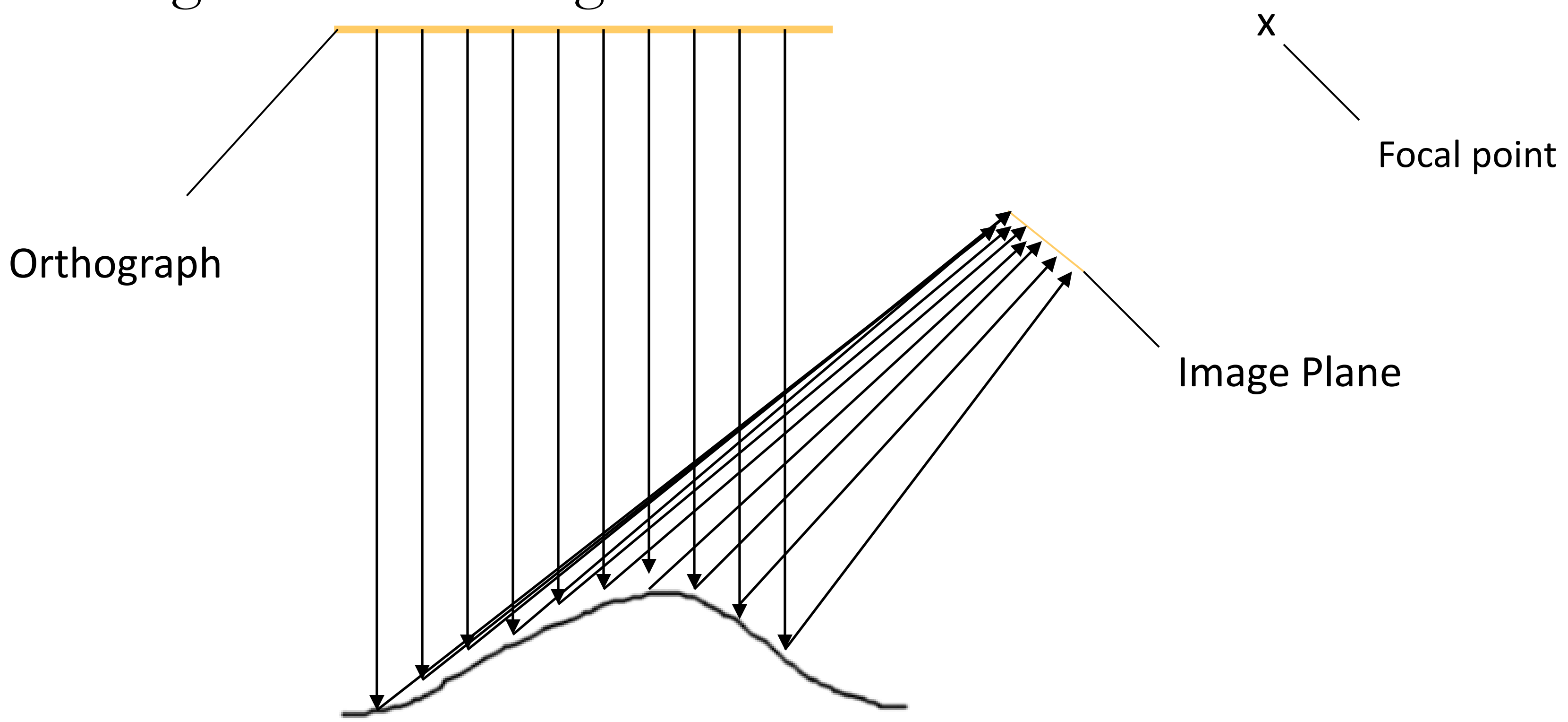
Rectification

Creating the OrthoImage



*ortho*Rectification

Creating the OrthoImage



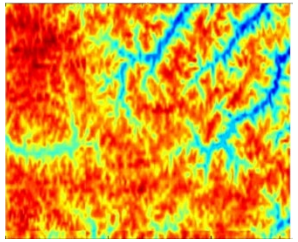
*ortho*Rectification

Sensor File + Mission Images + Reference Eniv. = Orthorectification

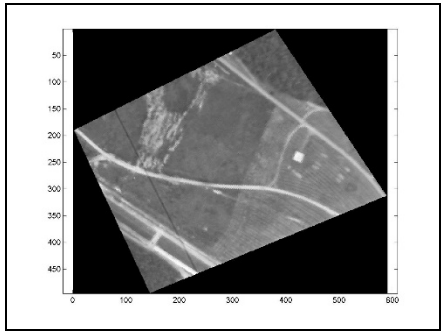
OTTER	system_id
TV	sensor_type
0001	serial_number
9.400008152666640300e+08	image_time
3.813193746469612200e+01	vehicle_latitude
-7.734523185193877700e+01	vehicle_longitude
9.949658409987658800e+02	vehicle_height
9.995171174441039900e-01	vehicle_pitch
1.701626418113209000e+00	vehicle_roll
1.207010551753029400e+02	vehicle_heading
1.658968732990974800e-02	camera_focal_length
-5.361314389557259100e+01	camera_elevation
-7.232969433546705000e+00	camera_scan_angle
480	number_image_lines
640	number_image_samples
1.000000000000000000e+00	u_axis_scale_factor
1.000000000000000000e+00	v_axis_scale_factor
3.812941913638164900e+01	ground_ref_pt_lat
-7.733710187470967400e+01	ground_ref_pt_lon
-1.418041566542428400e-09	ground_ref_pt_hgt
1.732050807568877200e+02	position_error
1.000000000000000000e+02	vehicle_lat_sigma
1.000000000000000000e+02	vehicle_lon_sigma
1.000000000000000000e+02	vehicle_hgt_sigma
5.000000000000000000e-01	vehicle_heading_sigma
5.000000000000000000e-01	vehicle_pitch_sigma
5.000000000000000000e-01	vehicle_roll_sigma
6.186380463322359200e-04	focal_length_sigma
5.000000000000000000e-01	gimbal_azimuth_sigma
5.000000000000000000e-01	gimbal_elevation_sigma

+

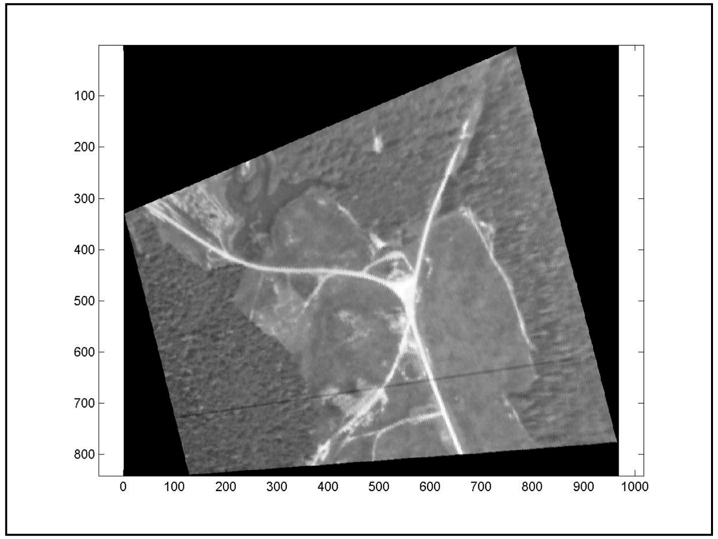
+



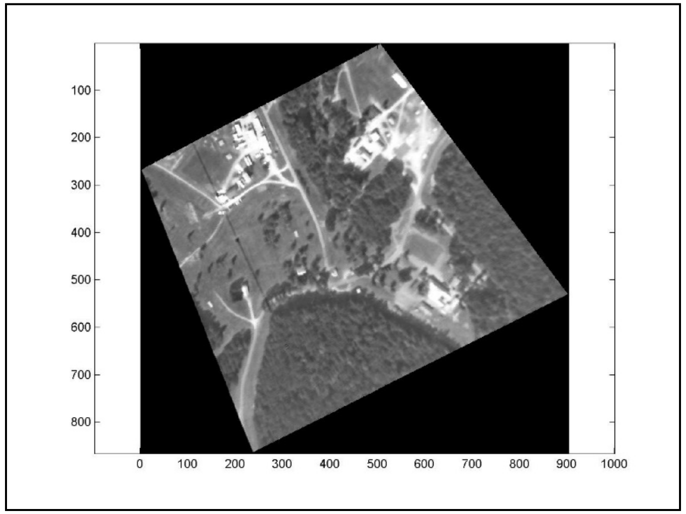
=



*ortho*Rectification



*ortho*Rectification



Reference Image
(DOQ)

Contents

- Pixel-Wise Geo-localization
  - Geodetic Alignment of Aerial Video Frames
- **Image-Based Geo-Localization**
  - Same View (Street-View to Street-View)
    - Generalized Maximum Clique (PAMI, 2014)
    - Constraint Dominant Sets (PAMI, 2017)
  - Cross-View Geo-Localization
    - Bird's Eye-View to Street View (CVPR, 2017)
    - Aerial to Ground View (ICCV, 2019)

Geo-localization Using Image Matching

- No Meta-Data (Telemetry)
- No Sensor Model
- No Geodetically Accurate Reference Image
  - Only Geo-tagged images
- Image level coarse geo-localization

"Where Am I?"

- Problem:

  Image Localization

Input

Output

Mere Visual Information(Images)   Location in Terms of λ (Lon.) and φ (Lat.)

“Where Am I?”

- Problem:

  Image Localization

Input

Output

Mere Visual Information(Images)

Location in Terms of λ (Lon.) and φ (Lat.)

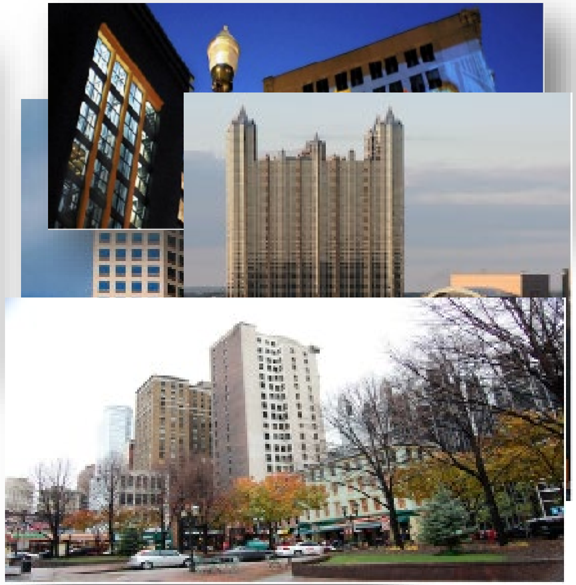
φ=40.4419, λ=-79.9986

Google Maps Street View Dataset

- Reference Set: GSV ~100k
- Test Set:
  - 521 GPS- Tagged from **Pittsburgh, PA and Orlando, FL.**
  - Downloaded From Flickr, Panoramio, Picasa, etc.

picasa.google.com

Panoramio.com



Flickr.com

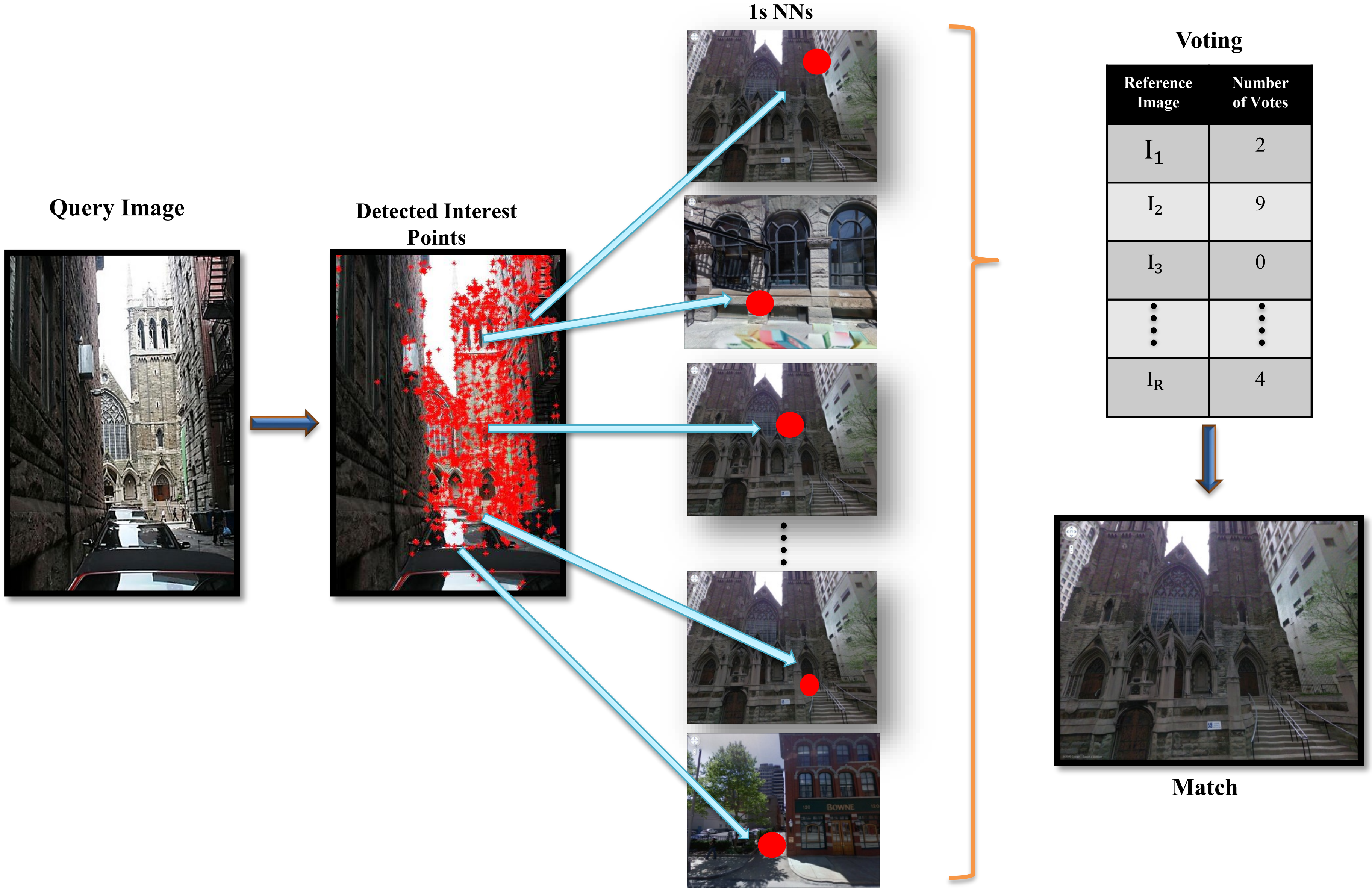
Image Geo-localization Based on Multiple Nearest Neighbor Feature Matching Using Generalized Graphs.

Amir Zamir and Mubarak Shah

In ***T-PAMI***, 2014.

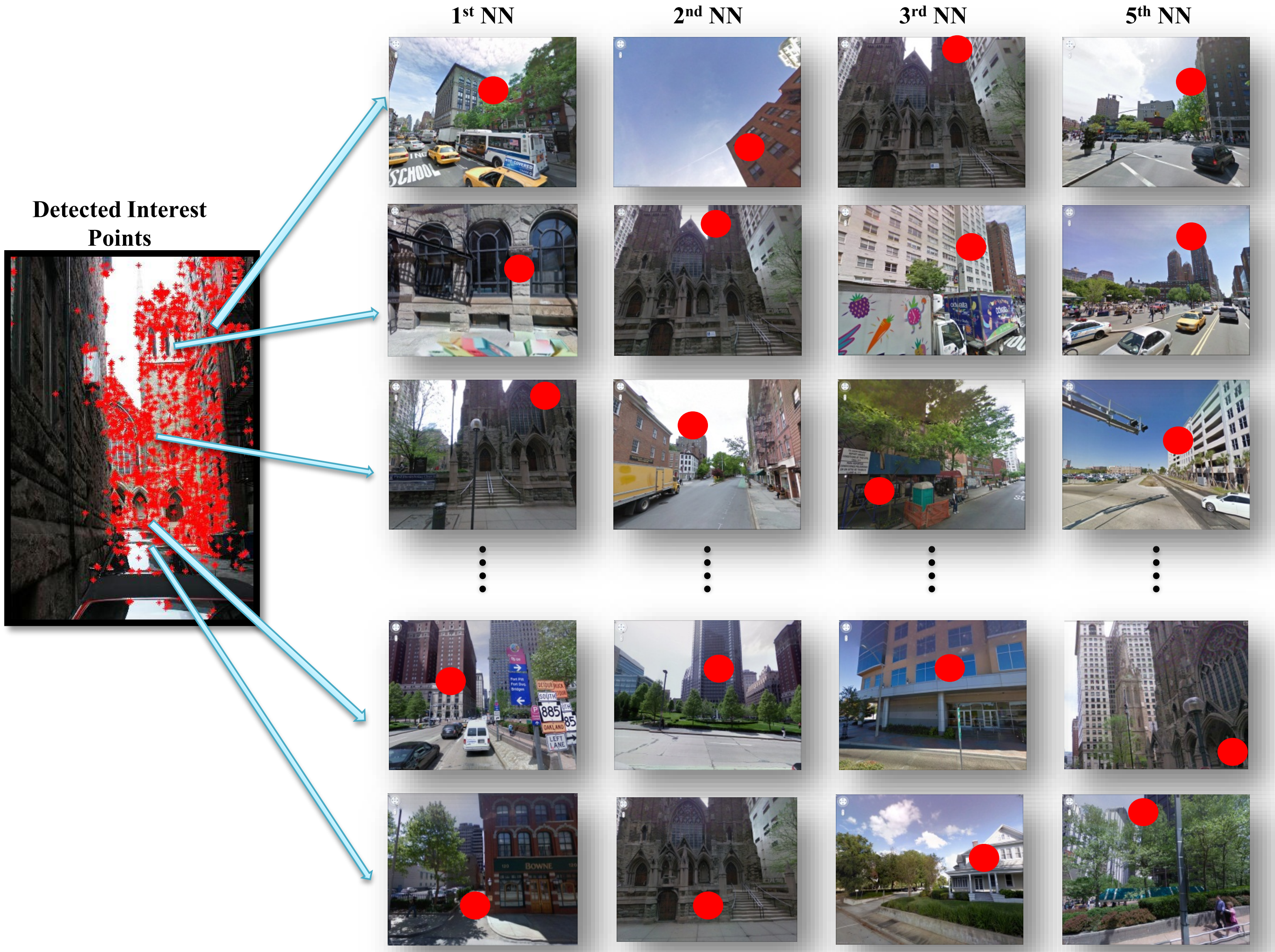
Asst Professor; EPFL

I: Localization Recognition Using Local Features

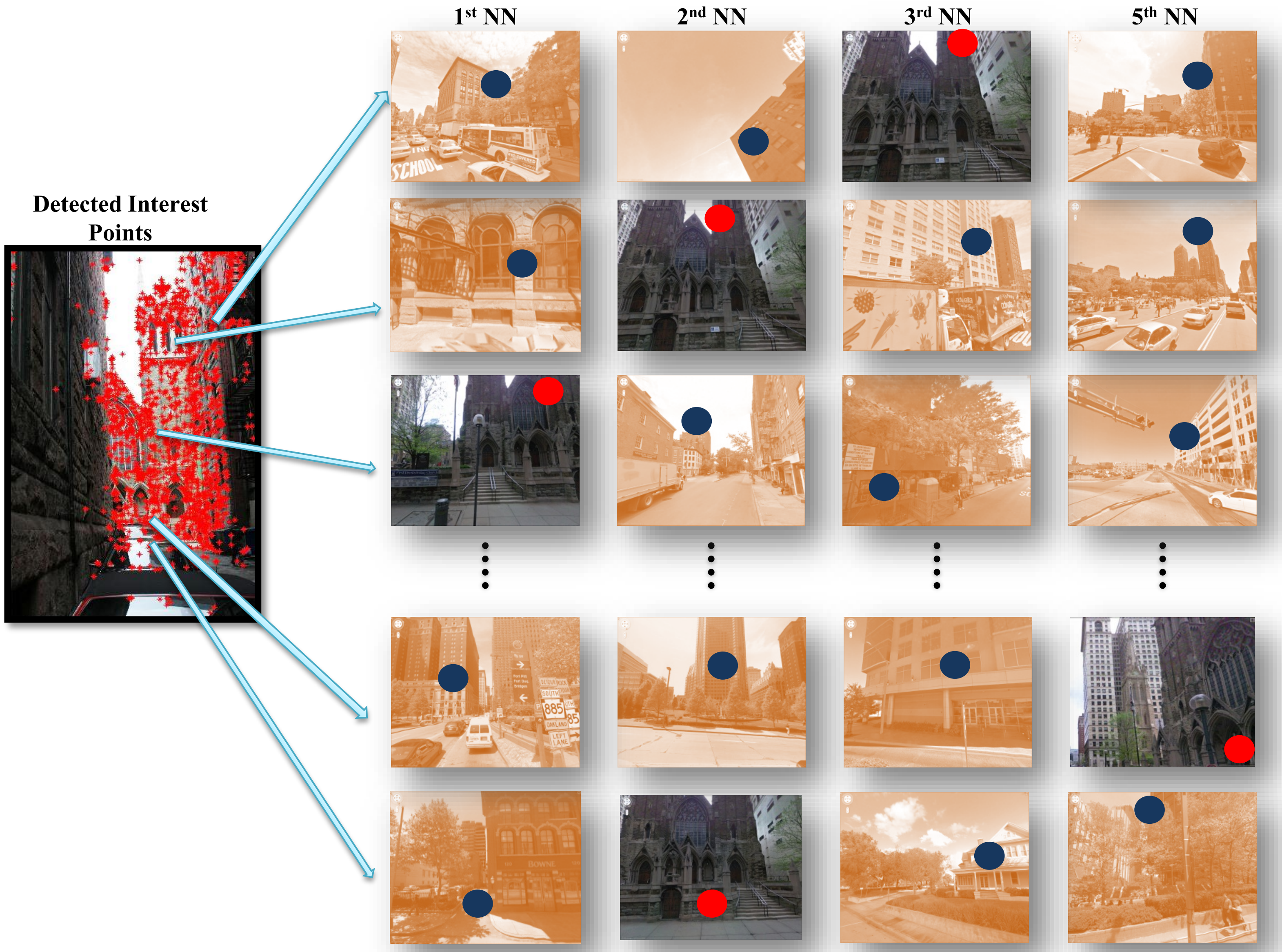


Reference Image	Number of Votes
I_1	2
I_2	9
I_3	0
⋮	⋮
I_R	4

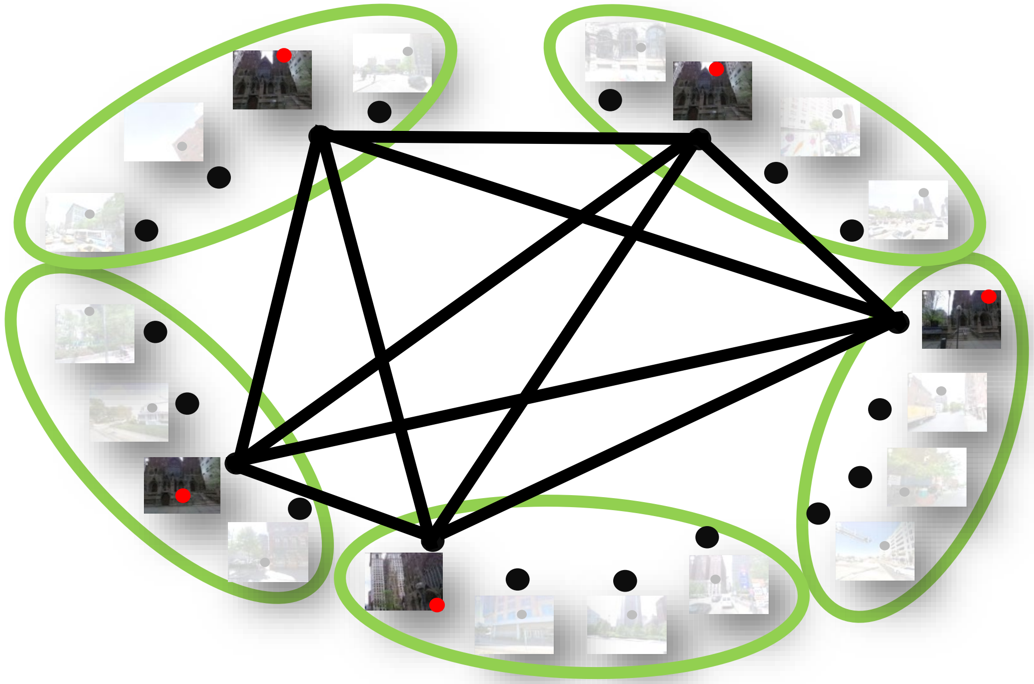
Using Multiple Nearest Neighbors



Using Multiple Nearest Neighbors



Generalized Minimum Clique



- Subset of NNs with maximum agreement in local and global features

Geo-localization Results

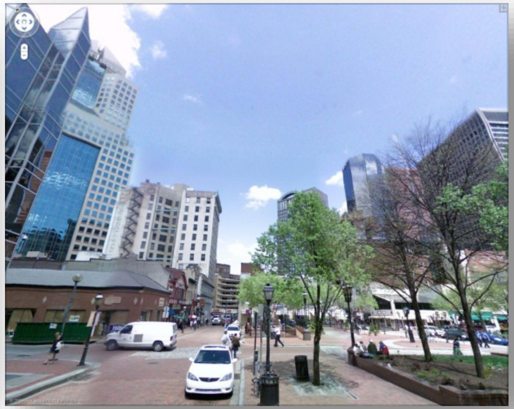
Query

Match – Error: 7.6 m

Query

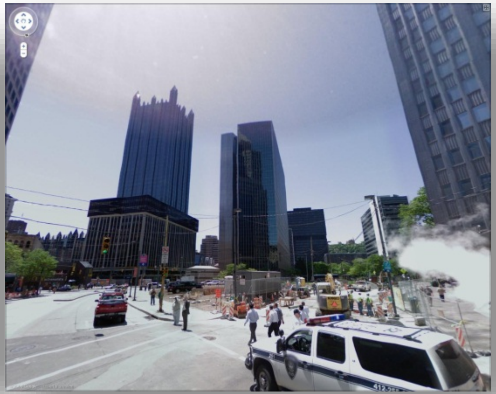
Match – Error: 6.9 m

Query



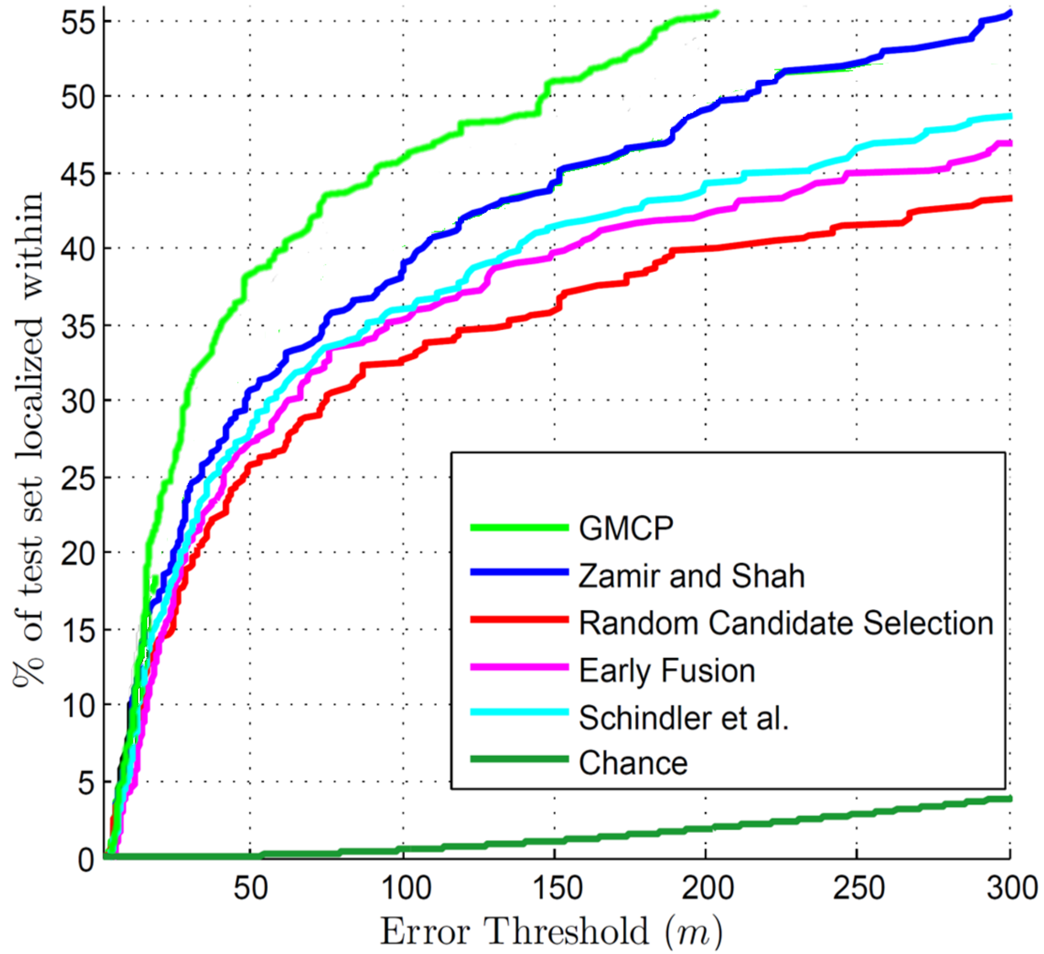
Match – Error: 308.1 m

Query



Match – Error: 59.3 m

Geo-localization Results



Limitations

- GMCP selects exactly one NN per query feature; sensitive to outliers
- A very simple voting scheme
- GMCP is a binary-variable NP hard problem

Image Geo-Localization Using Constraint Dominant Sets

Eyasu Mequaanint, Qualcom

Yonatan Tariku. , Qualcom

Haroon Idrees, RetailNext

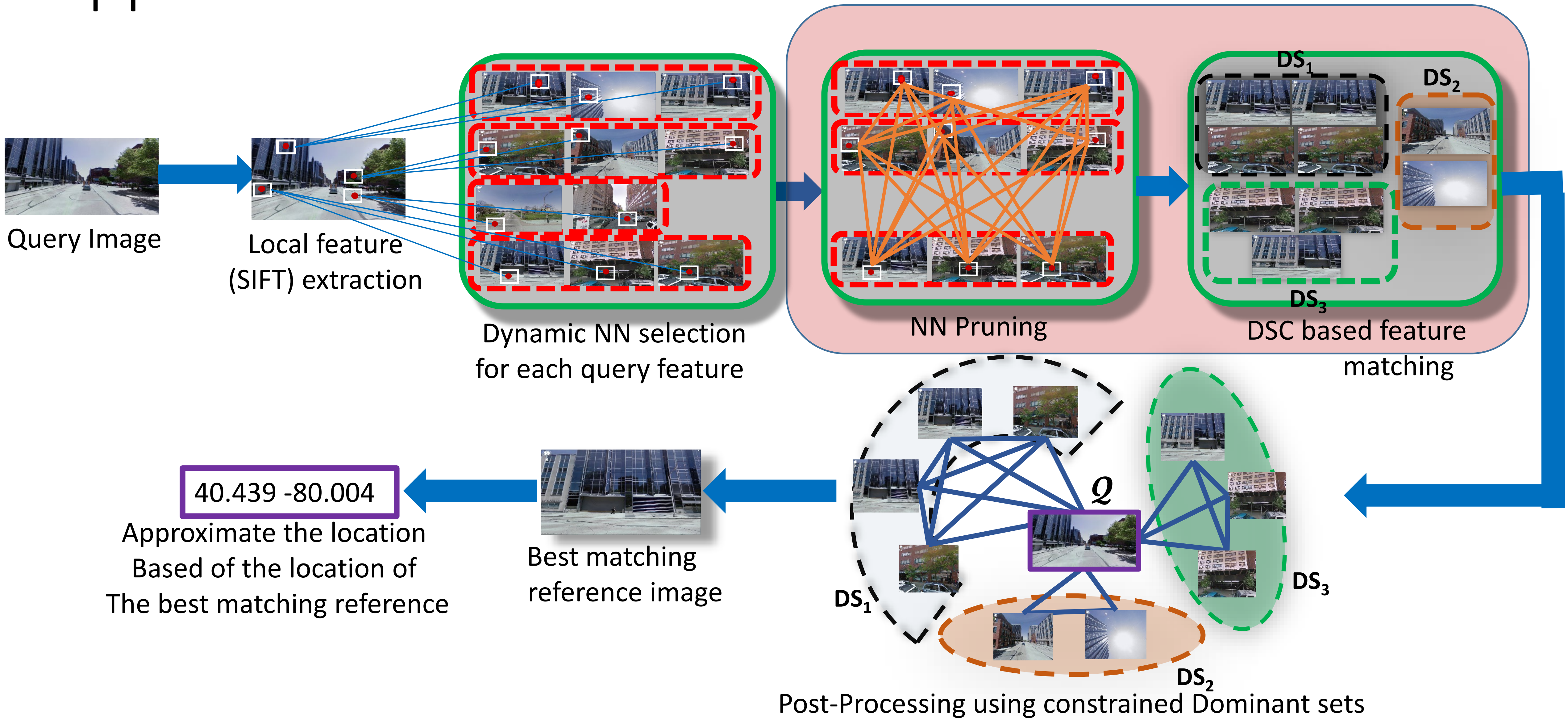
Eyasu Zemene,Yonatan Tariku Tesfaye, Haroon Idrees, Andrea Prati, Marcello Pelillo, and Mubarak Shah

T-PAMI, 2017.

Goal

- Fast
- Accurate
- Handle outliers
- Scalable to large scale

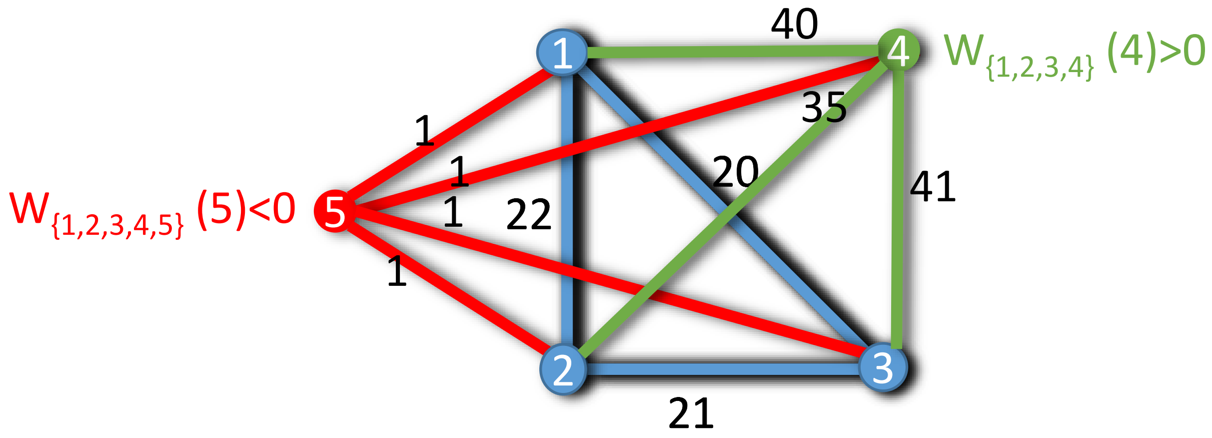
Approach


Query Image
Local feature
(SIFT) extraction
Dynamic NN selection
for each query feature
NN Pruning
DS_1
DS_2
DS_3
DSC based feature
matching
Q
DS_1
DS_3
DS_2
Post-Processing using constrained Dominant sets
Best matching
reference image
40.439 -80.004
Approximate the location
Based of the location of
The best matching reference

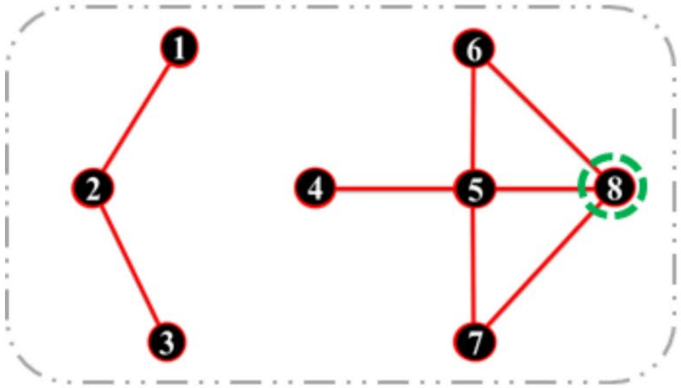
Dominant Sets

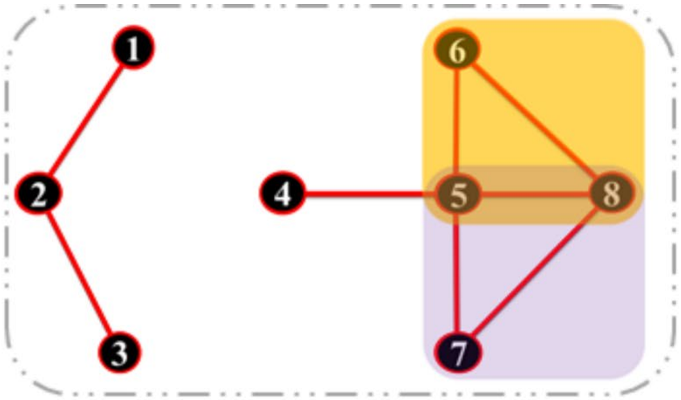
- Dominant set is an edge-weighted generalization of a clique
- Dominant set is a subset of vertices, which is
  - Coherent and
  - Compact

Dominant Sets



Constraint Dominant Sets





Dominant Sets

- Edge-weighted generalization of maximal cliques
- Given a (symmetric) affinity **A**, consider,

$$\begin{array}{ll} \text{maximize} & f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} \\ \text{subject to} & \mathbf{x} \in \Delta \end{array}$$

Where $\Delta = \{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \text{ and } x_i \geq 0, \text{ for all } i = 1, \ldots, n\}$

- If **x** is a local maximizer, its support, $\sigma(\mathbf{x})$, is a dominant set
- DS's caputer both *internal* and *external* coherence conditions for a cluster

M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 167–172, 2007.

Constrained Dominant Sets (CDS)

Given a query $\mathcal{Q} \subseteq \mathrm{V}$ and a parameter $\alpha > 0$, define the following parameterized family of quadratic program:

$$maximize\ \ f_{\mathcal{Q}}^{\alpha} = x^T\,(A - \alpha I_{\mathcal{Q}}\,)x \qquad (1)$$
$$\text{Subject to } x \in \Delta$$

Where $I_{\mathcal{Q}}$ is the diagonal matrix whose **diagonal elements are set to 1** in correspondence to the vertices contained in $\mathrm{V} \setminus \mathcal{Q}$ and to 0 otherwise.

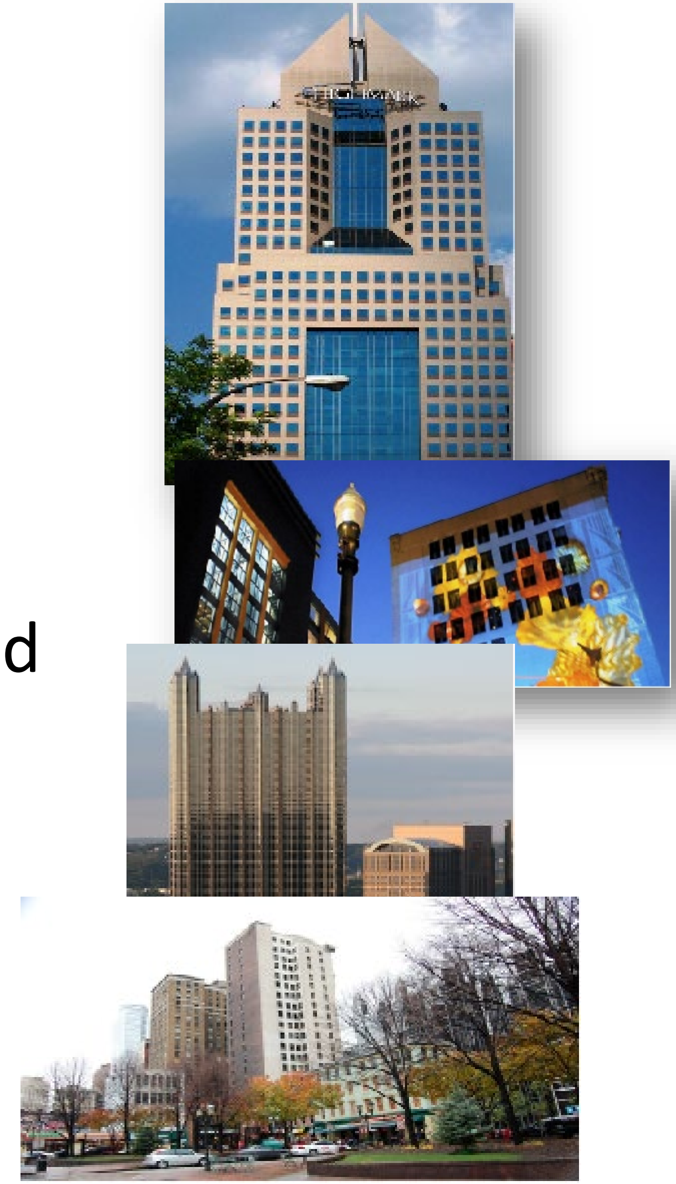
Experimental Results

Dataset I: Orlando & Pittsburgh:

- Reference images:
  - 102K Google street view images from **Pittsburgh, PA** and **Orlando, FL**
- Test Set:
  - 644 GPS-Tagged unconstrained images
  - Downloaded From Flickr, Panoramio, Picasa, …

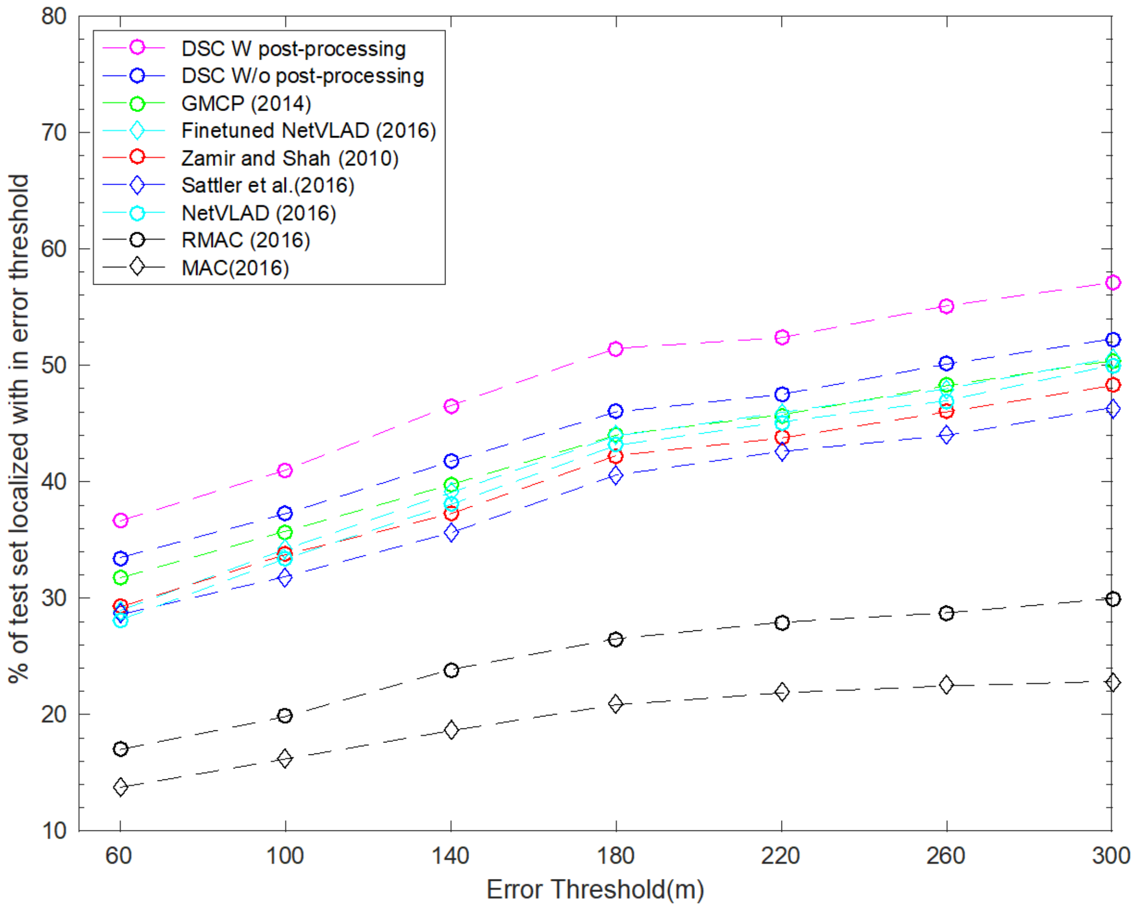
Dataset II: *WorldCities* Dataset: (NEW)*

- **Reference images** (300K Google street view images):
- 14 different cities from different parts of the world:
  - USA: Los Angeles, Phoenix, Houston, San Diego, Las Vegas, Dallas, Chicago
  - Australia: Sydney and Melbourne
  - Europe: Amsterdam, Frankfurt, Rome, Milan and Paris
  - **Test Set**
    - 500 GPS-Tagged unconstrained images
    - Downloaded From Flickr, Panoramio, Picasa…



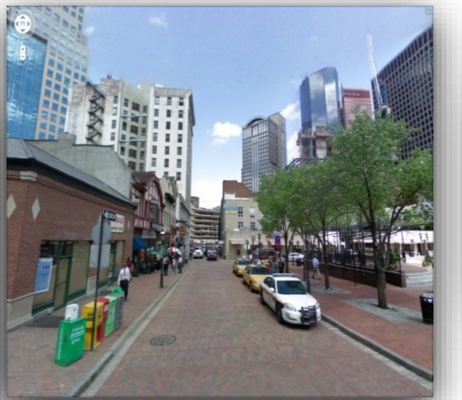
Overall Results

- Dataset 2: WorldCities (14 different cities from Europa, North America, Australia)



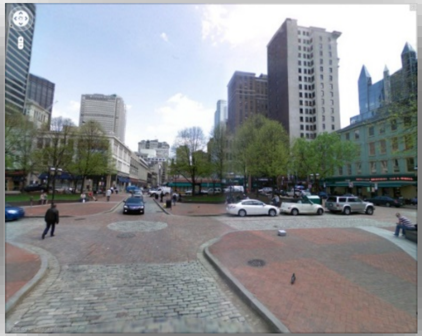
Qualitative Image Localization Results

Query



Match – Error: 70.01 m

Query



Match – Error: 5.4 m

Query

Match – Error: 10.4 m

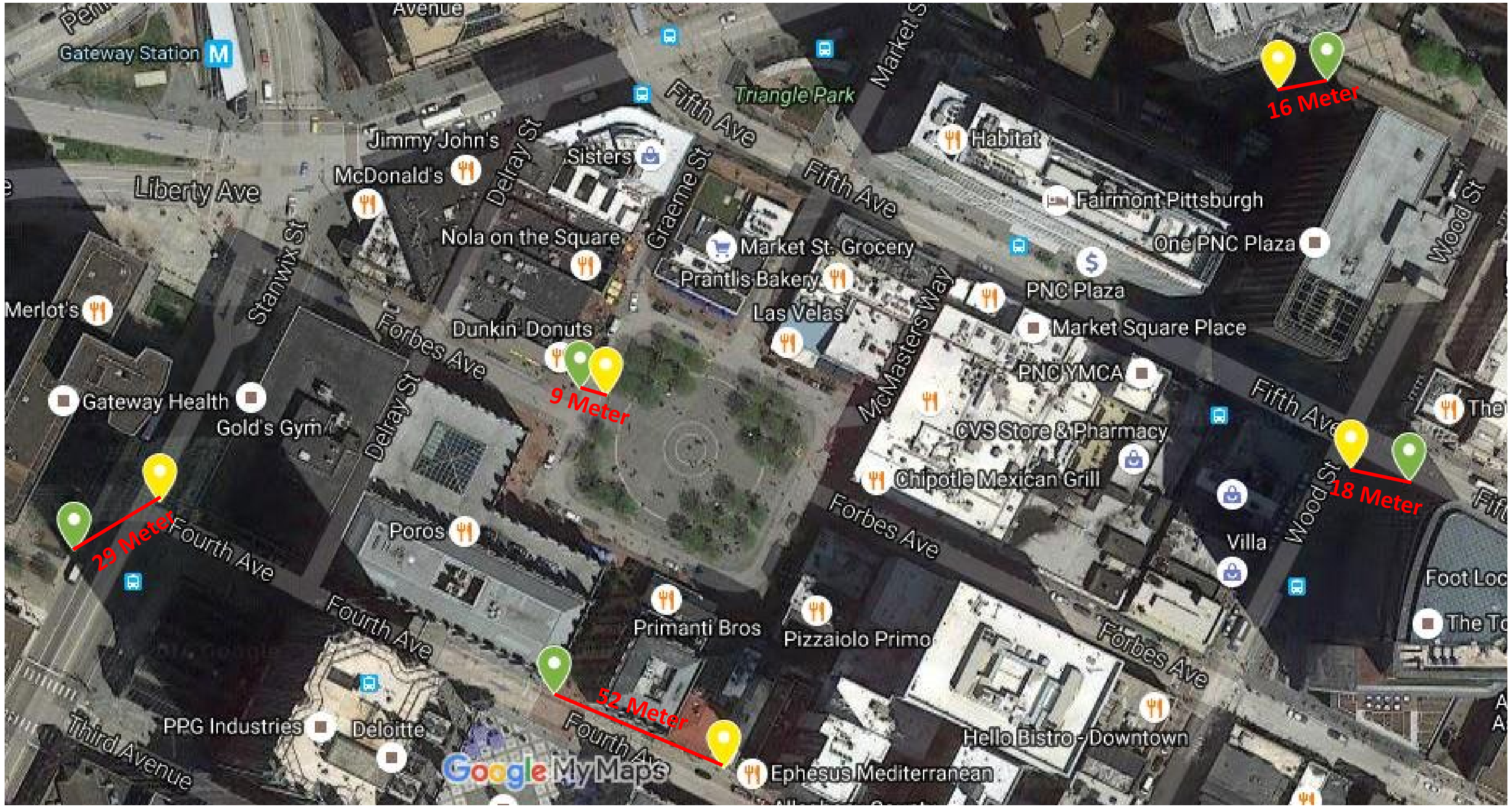
Query

Match – Error: 7.5 m

Query

Match – Error: 62.7 m

Qualitative Image Localization Results



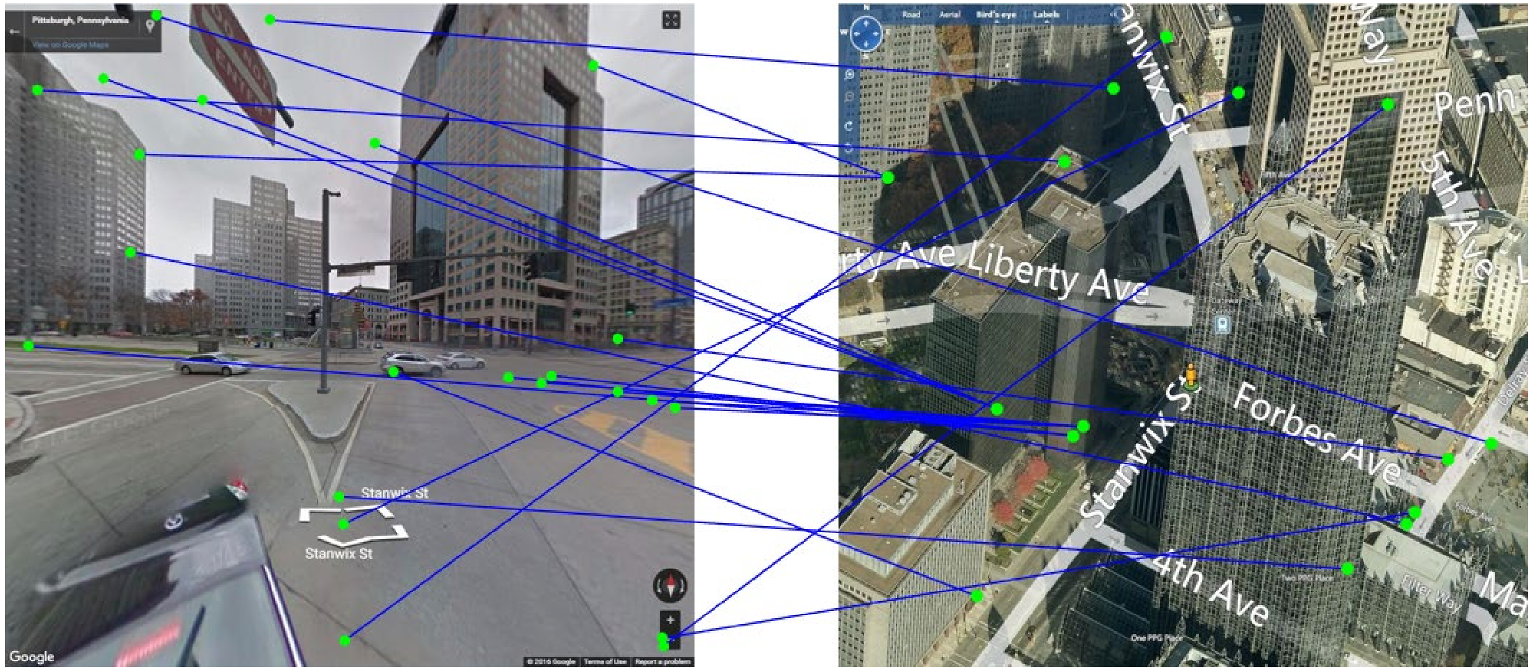
Contents

- Pixel-Wise Geo-localization
  - Geodetic Alignment of Aerial Video Frames
- Image-Based Geo-Localization
  - Same View (Street-View to Street-View)
    - Generalized Maximum Clique (PAMI, 2014)
    - Constraint Dominant Sets (PAMI, 2017)
  - **Cross-View Geo-Localization**
    - Bird's Eye-View to Street View (CVPR, 2017)
    - Aerial to Ground View (ICCV, 2019)

Cross-View Challenges

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

- Images from different viewpoints are visually different
- The images may be captured with different lighting
- The mapping from one viewpoint to the other may be complex
- Traditional features e.g. SIFT, HOG etc. may be very different



Retrieval Features

- Local Features (SIFT) (same view)
- **Buildings Features** (cross view)

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Cross-View Image Matching for Geo-localization in Urban Environments

Yicong Tian, Google

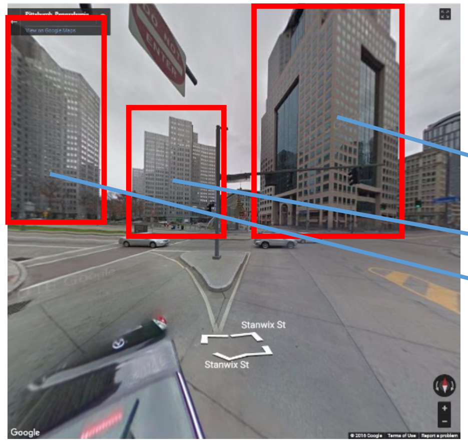
Yicong Tian, Chen Chen, Mubarak Shah
Center for Research in Computer Vision,
University of Central Florida

CVPR-2017

Chen Chen, UNC Charlotte

Proposed Approach

Query Image → Building Detection → Building Matching → Retrieve k NNs for each query building → Dominant sets selection → Geo-localization



Query image (street view)

Reference images (bird's eye view)

Data Collection

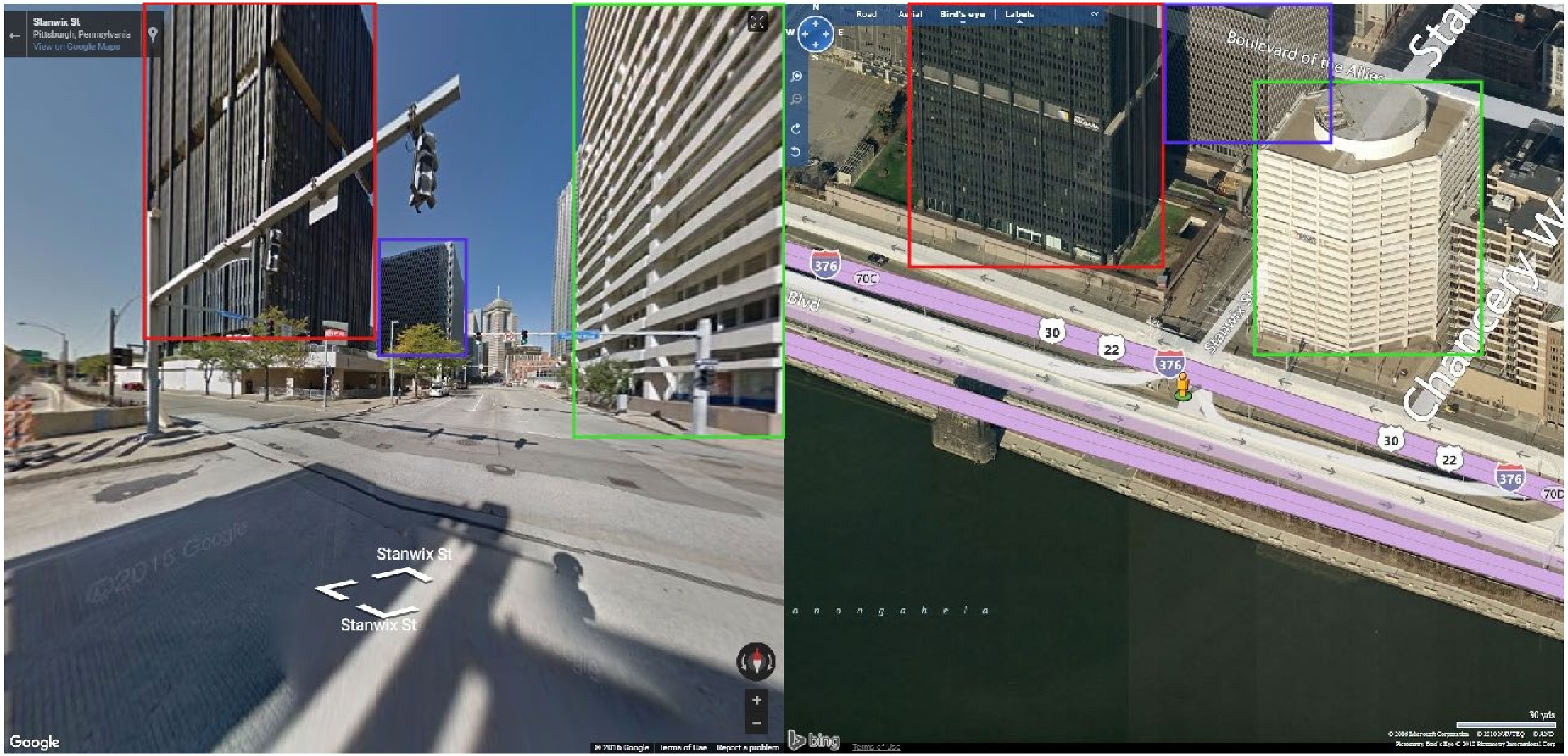
- A new dataset of street view and bird’s eye view image pairs
  - Pittsburg
  - Orlando
- Generated a list of GPS coordinates along streets
- For each GPS location
  - Four pairs with different headings
  - Utilized DualMaps
  - Automatically saved screenshots
- Annotations
  - By four undergraduates and high school students
  - ~300 hours work in total

Data Collection - Pittsburg

- 1,586 GPS locations

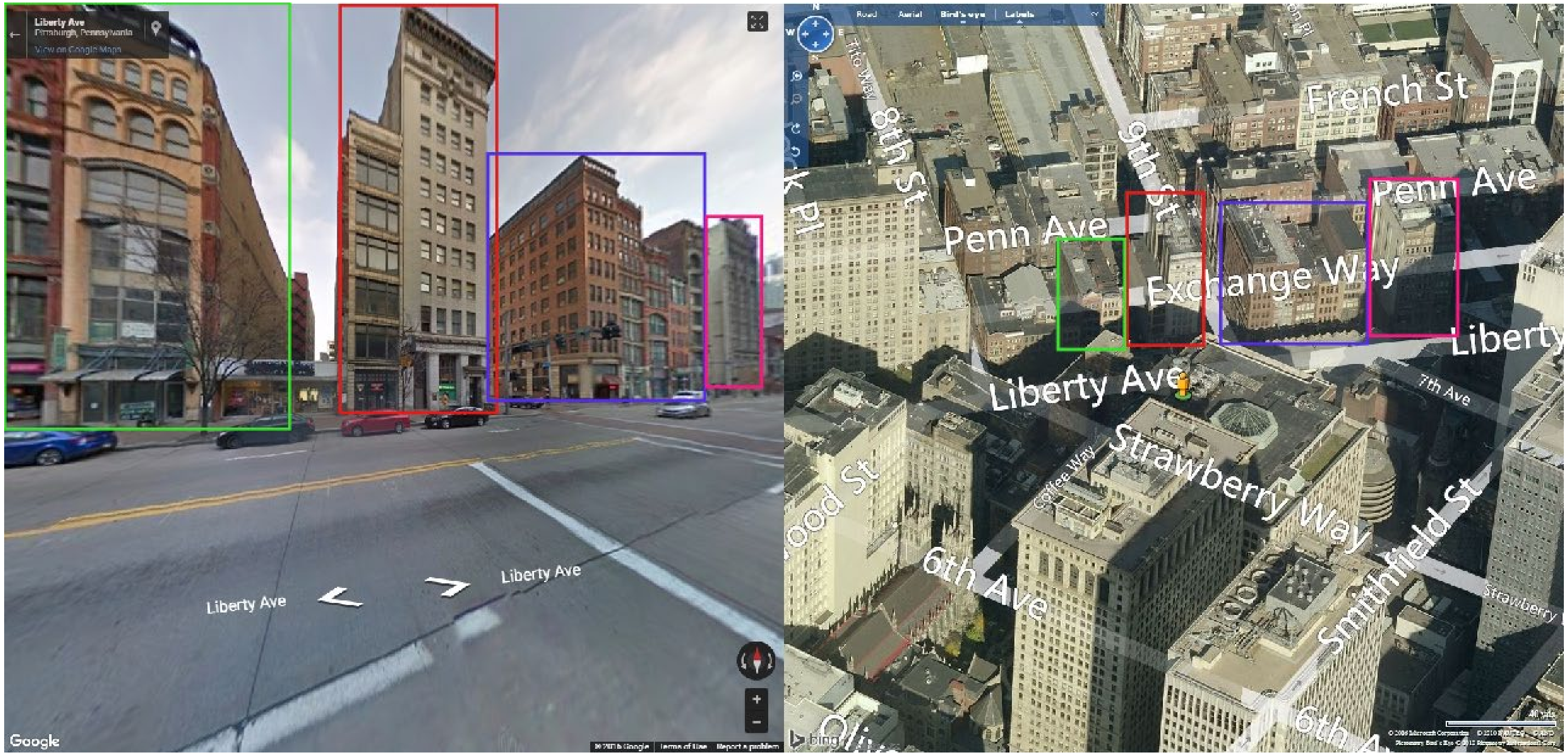
Data Collection - Pittsburg

- Example image pair with annotations



Data Collection - Pittsburg

- Example image pair with annotations

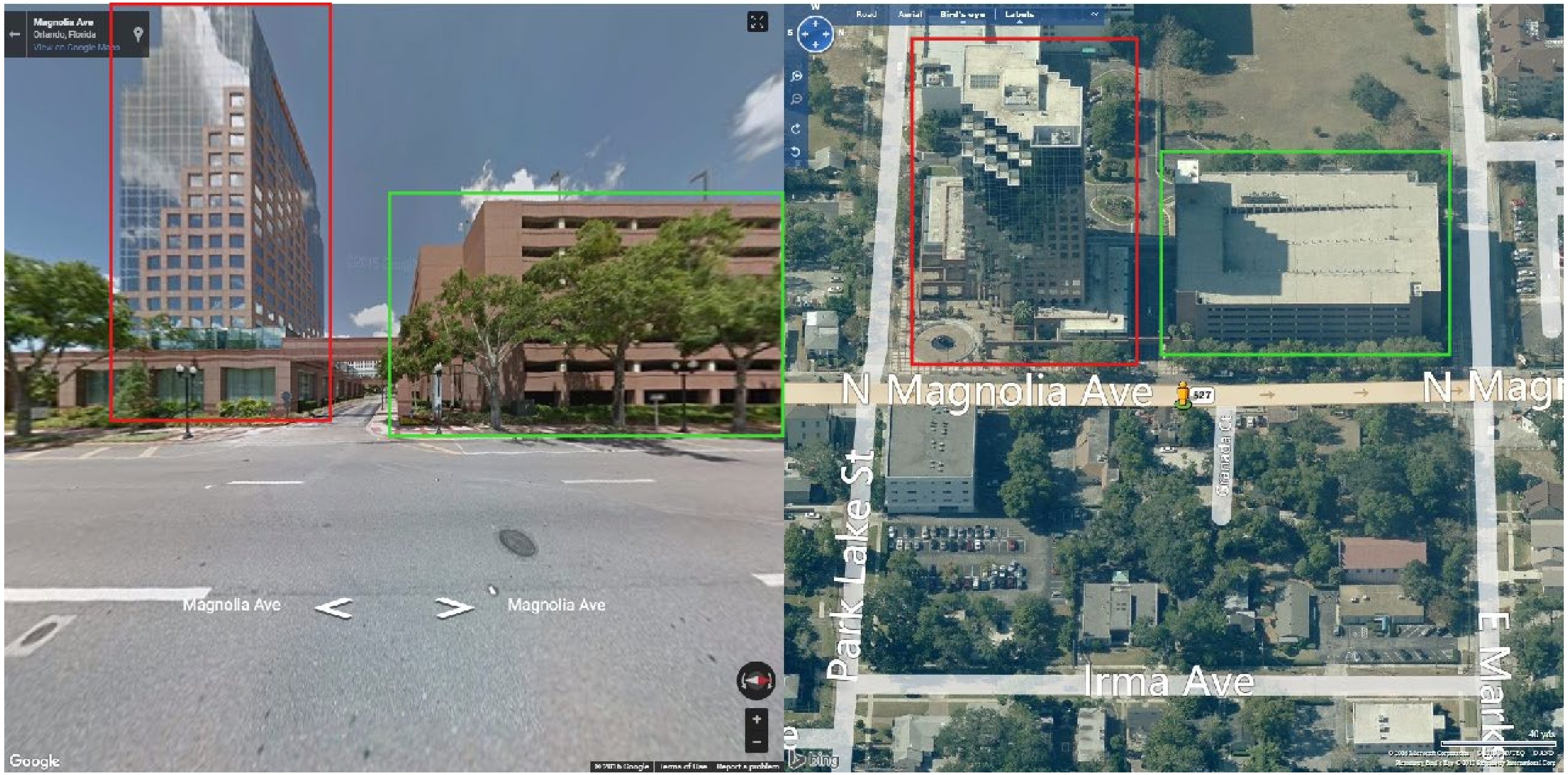


Data Collection - Orlando

- 1,324 GPS locations

Data Collection - Orlando

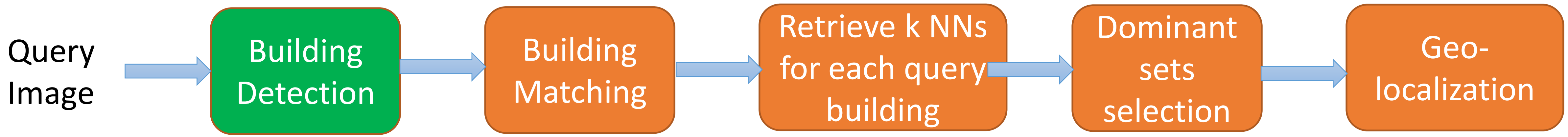
- Example image pair with annotations



Data Collection - Orlando

- Example image pair with annotations

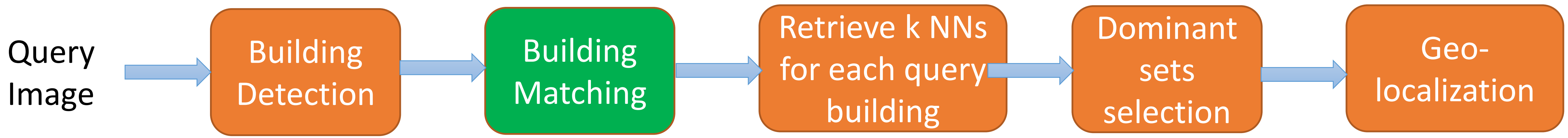
Proposed Approach



Building Detection (Faster R-CNN)

- Street view images
  - Training: 6,682 images
    - With 15,648 annotated boxes
  - Test: 1,903 images
- Bird’s eye view images
  - Training: 6,968 images
    - With 39,511 annotated boxes
  - Test: 1,916 images
- Each model takes 10 hours to train

Proposed Approach



Building Matching

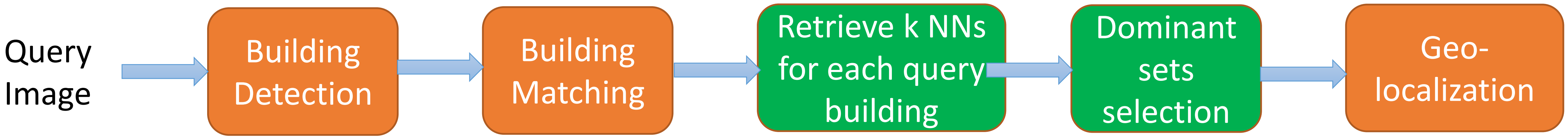
- Train a Siamese network
- In the learned feature space,
  - Matching image pairs are close to each other
  - Unmatched image pairs are far apart
- Contrastive loss

$$\mathcal{L}(x, y, l) = \frac{1}{2} l D^2 + \frac{1}{2}(1 - l) \max\left(0, (m - D^2)\right)$$

l: label (1 or 0)
D^2: square of Euclidean distance between features
m: margin

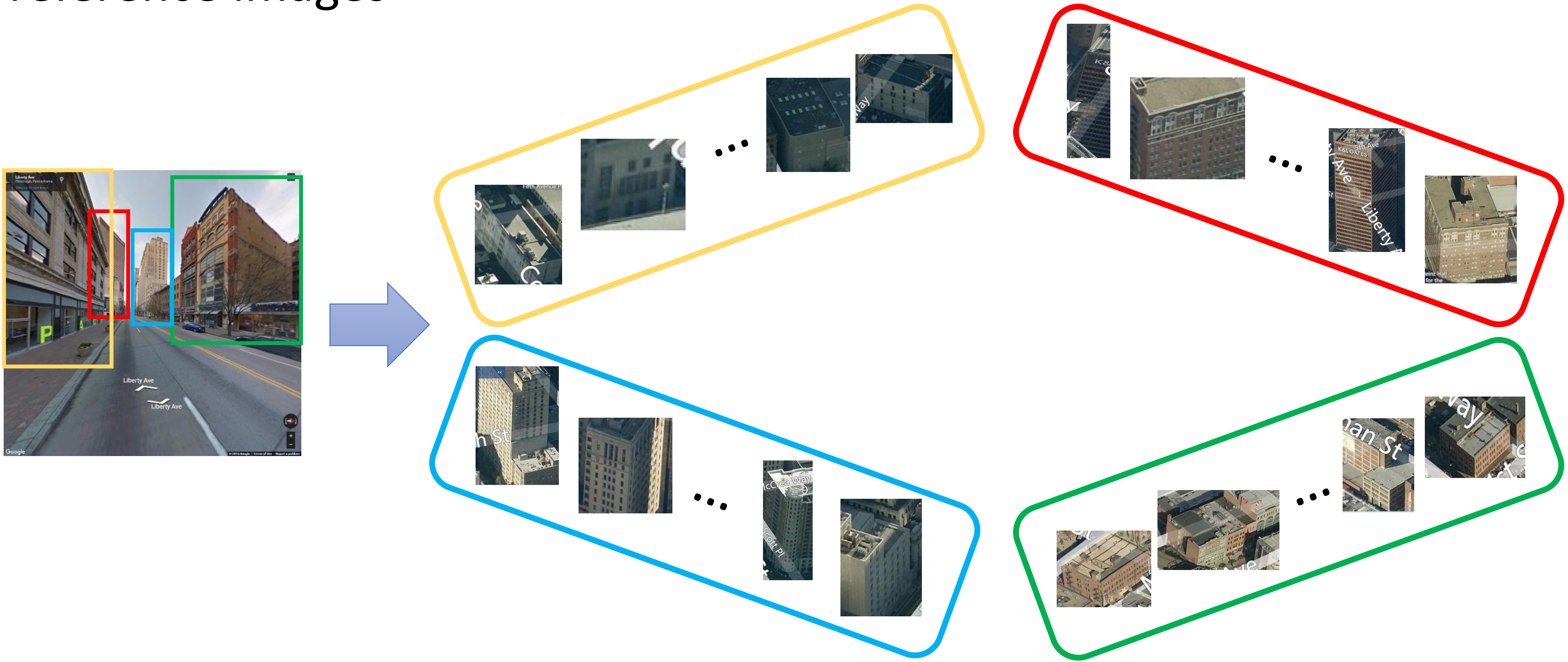
Proposed Approach

UCF Center for Research in Computer Vision



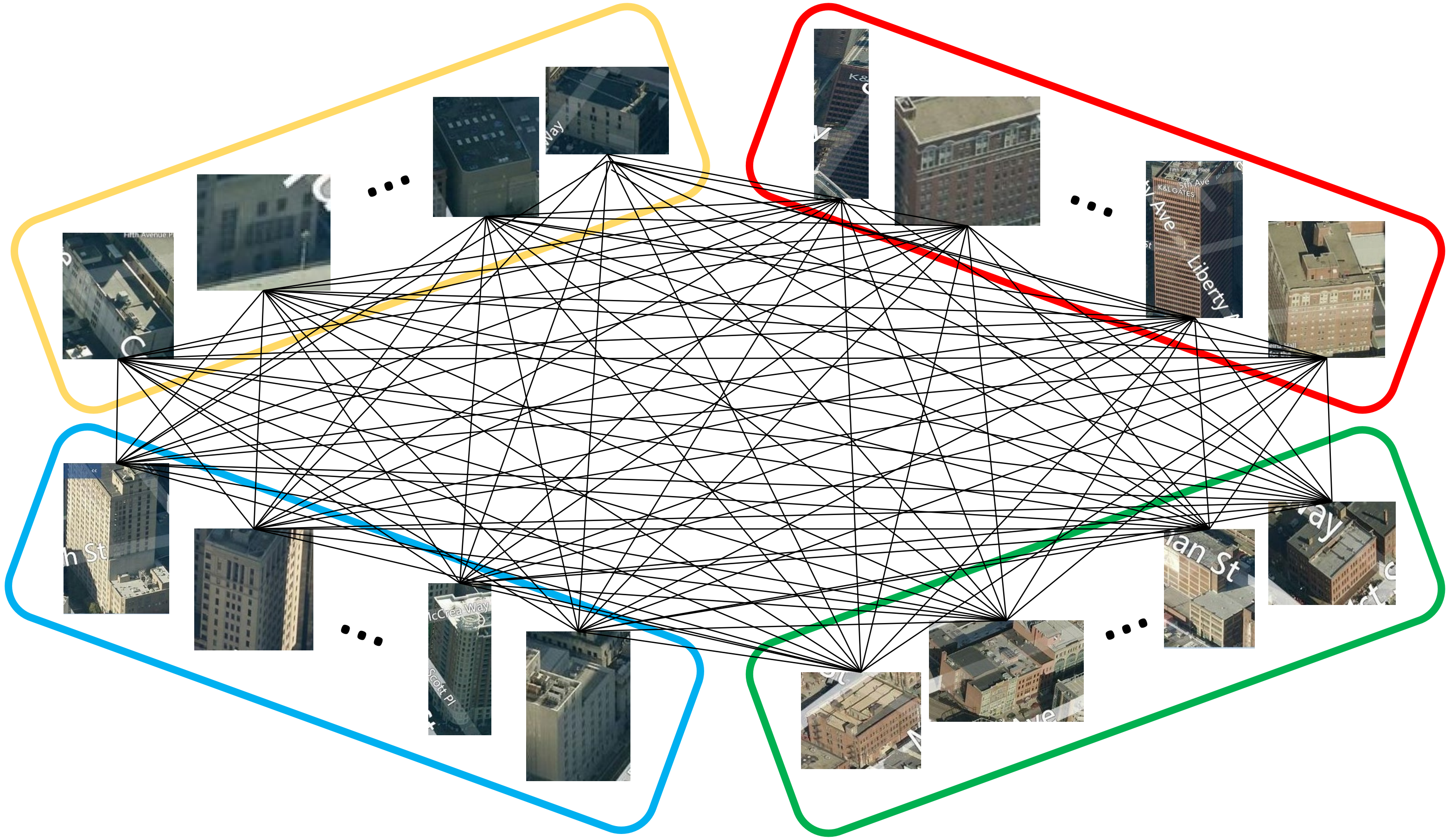
Geo-localization Using Dominant Sets

- For each building in the query image, select k nearest neighbors from reference images



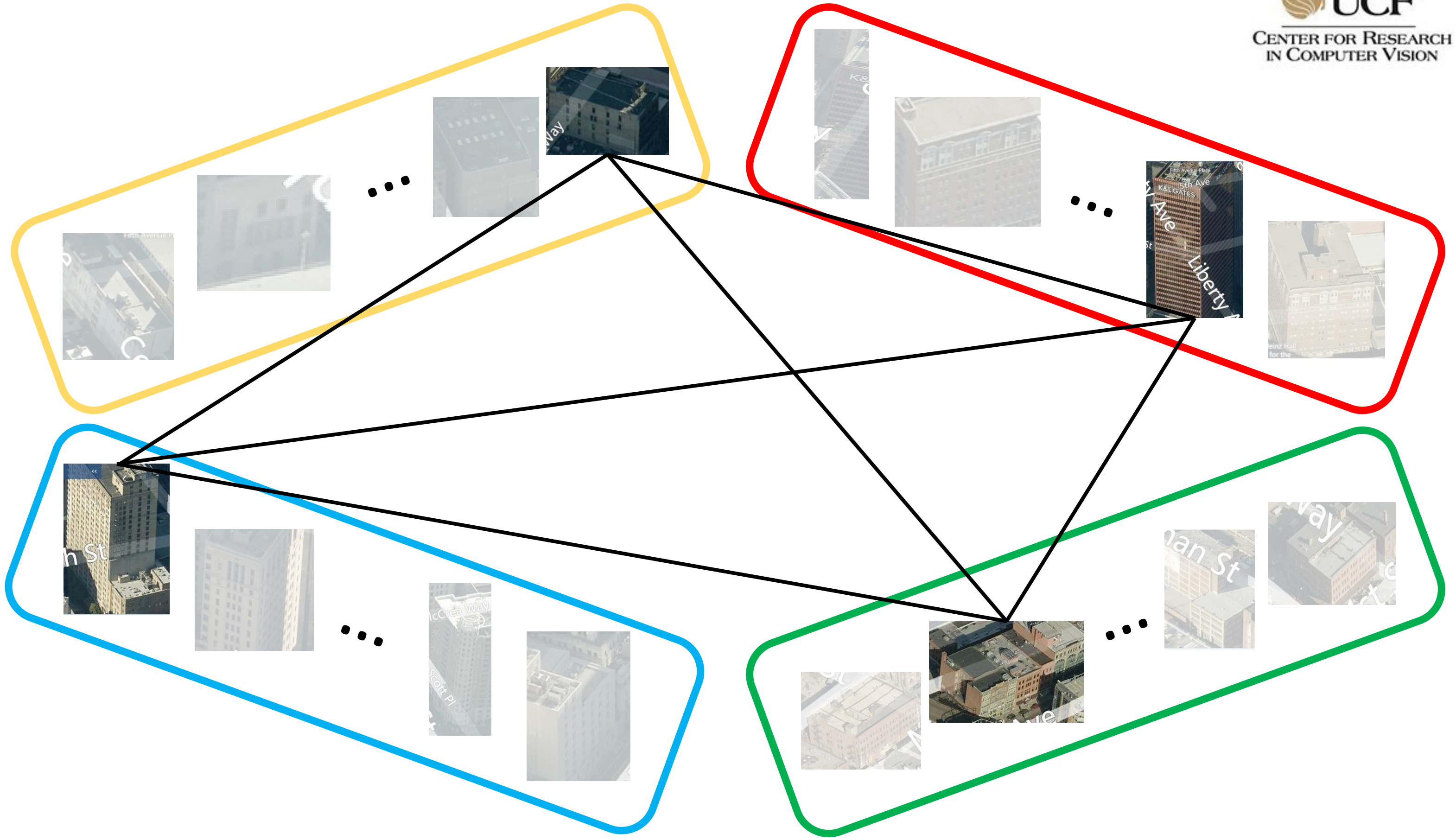
Geo-localization Using Dominant Sets

- Build a graph $G = (V, E, \omega)$ using selected reference buildings



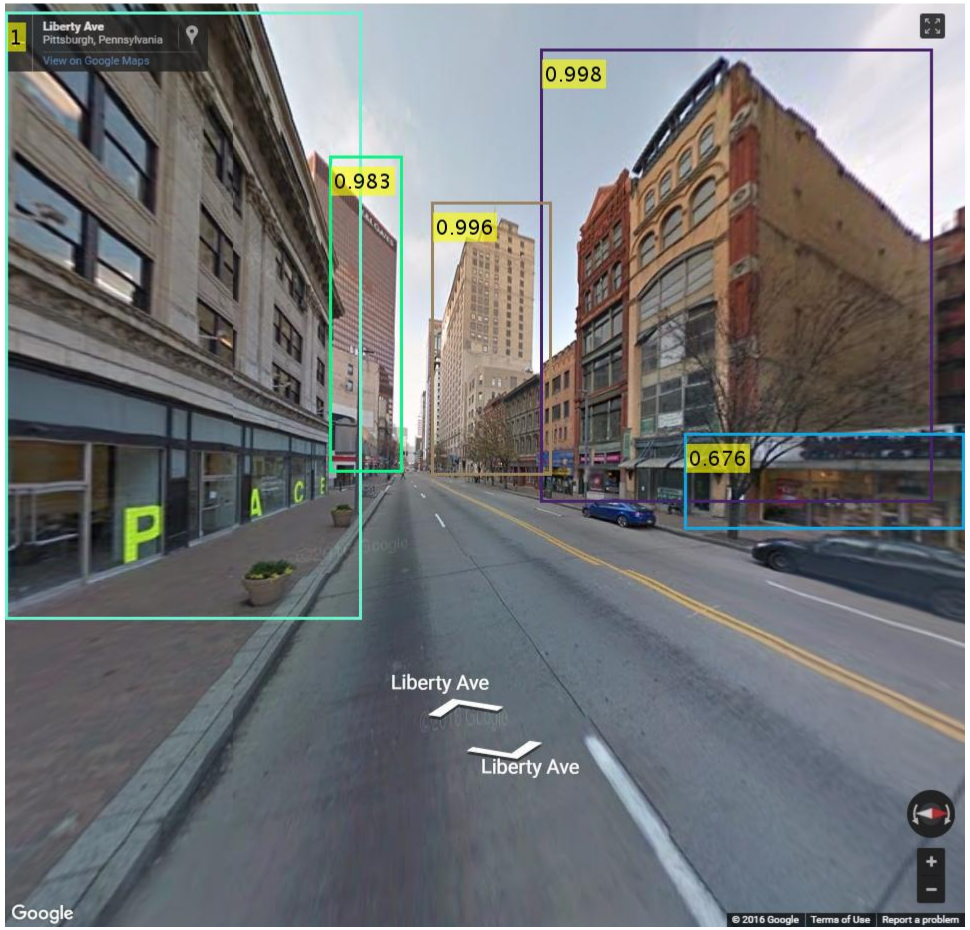
Geolocalization

- The nodes in dominant set form a coherent set
- At most one node is selected from each cluster



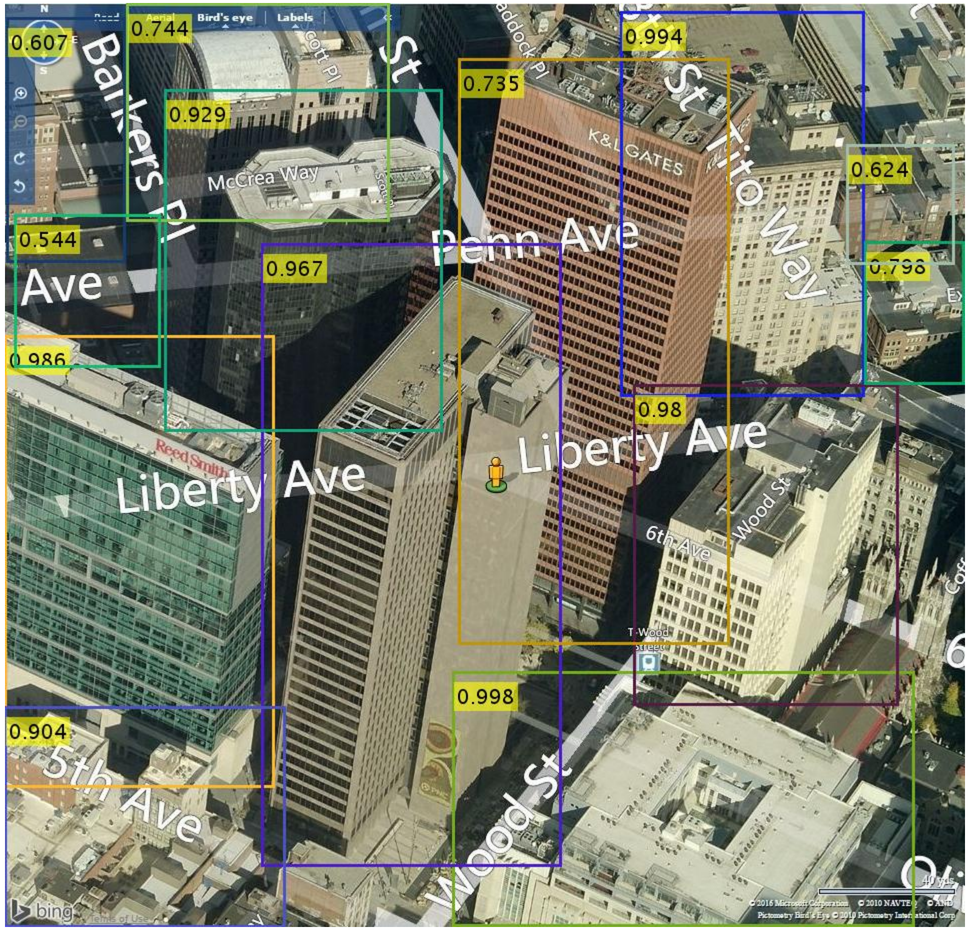
Experimental Results

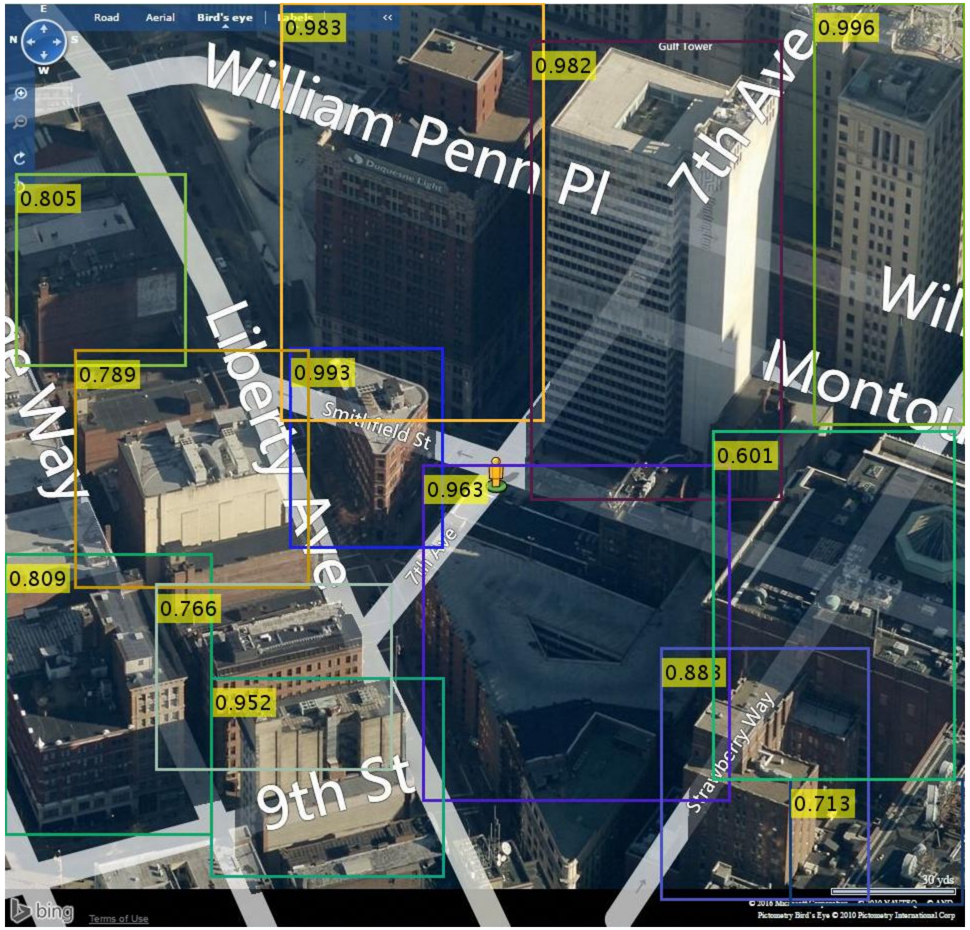
Building detection in street view images



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

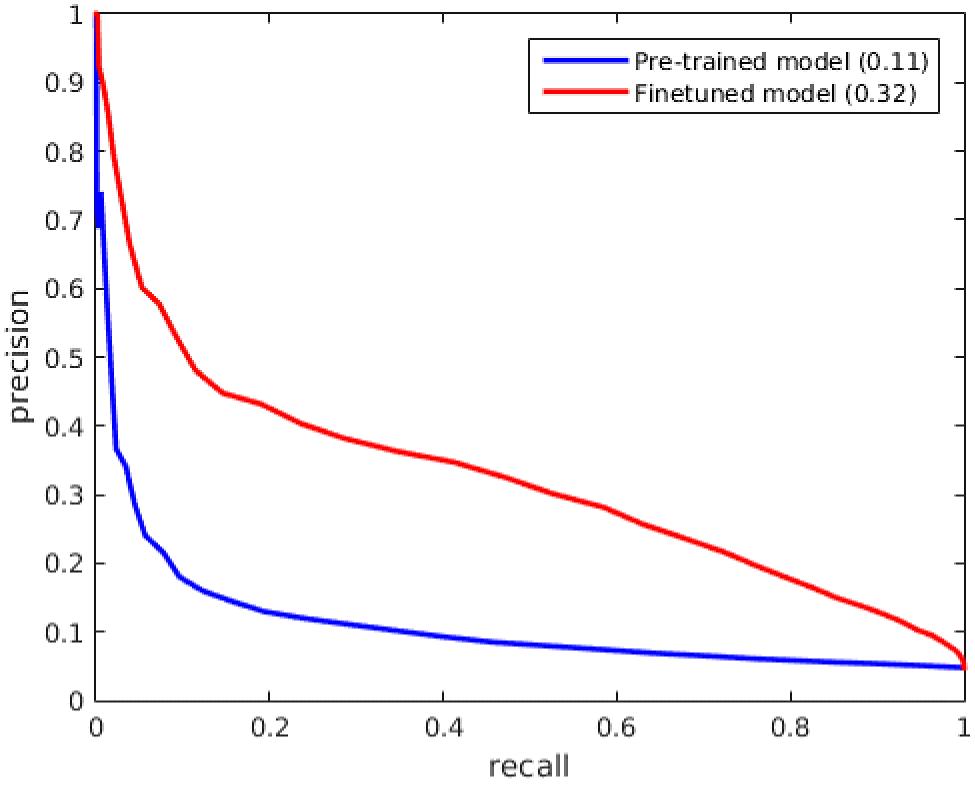
Building detection in bird's eye view images



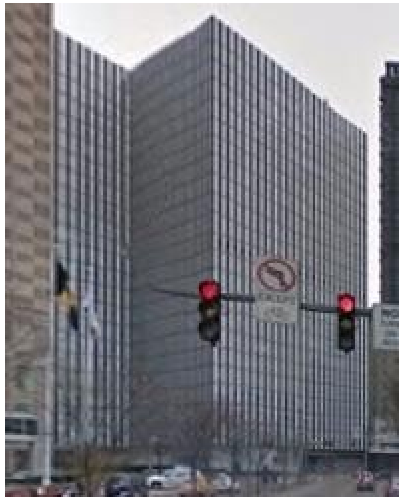


Building matching

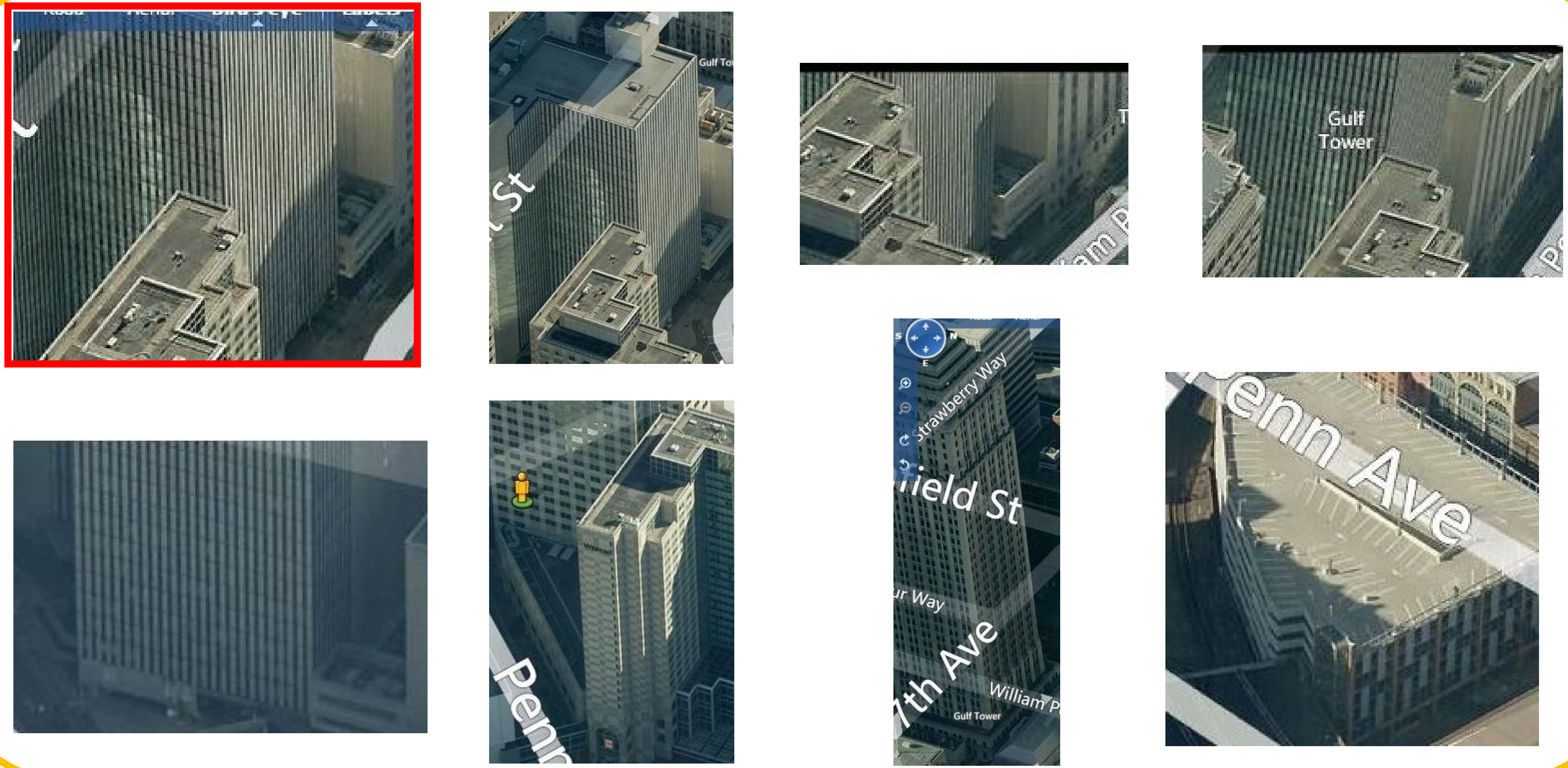
- Precision-recall curve on test image pairs



Building matching



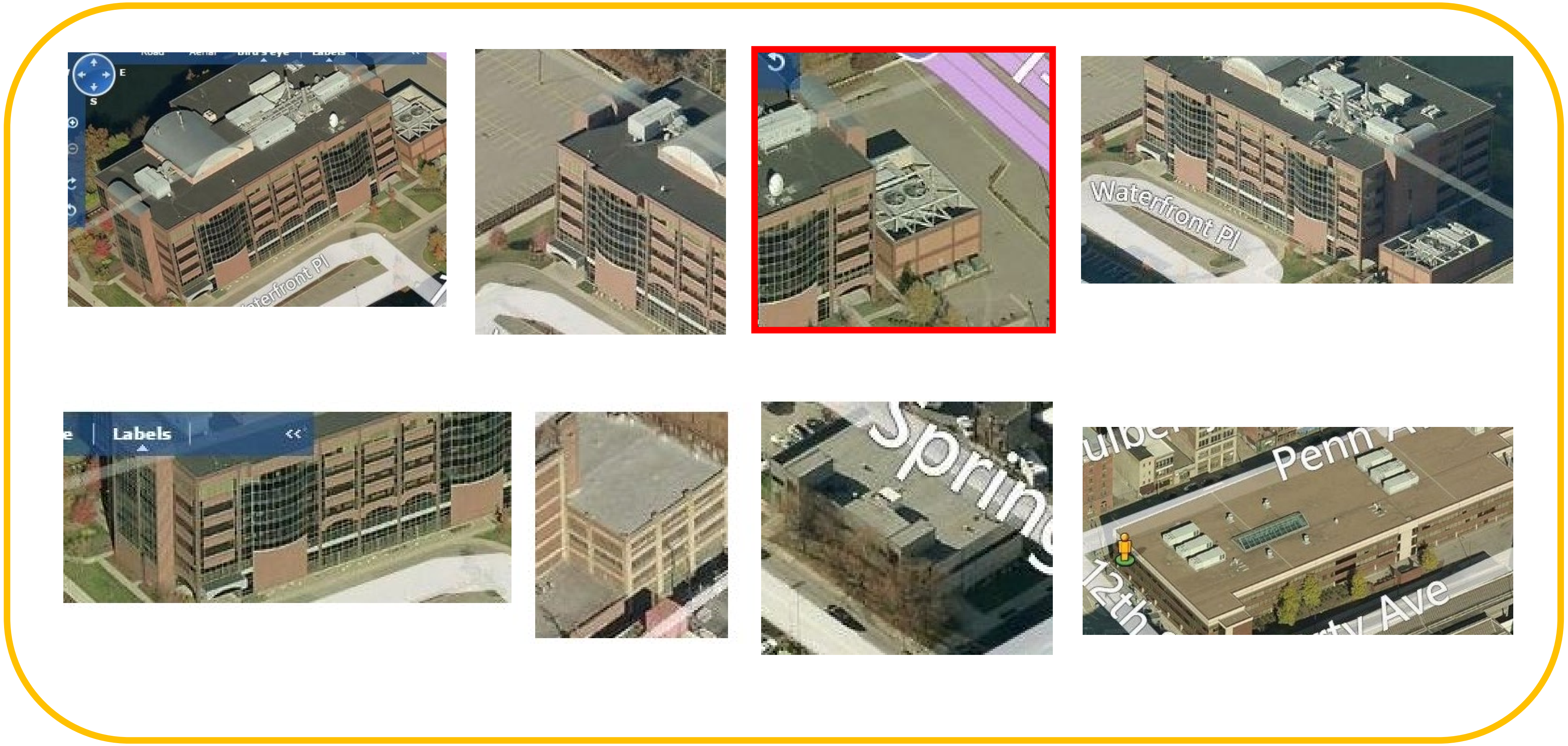
Query image
(street view)



From ~40,000 candidate image patches (Buildings)
(bird's eye view)

Building matching

Query image
(street view)

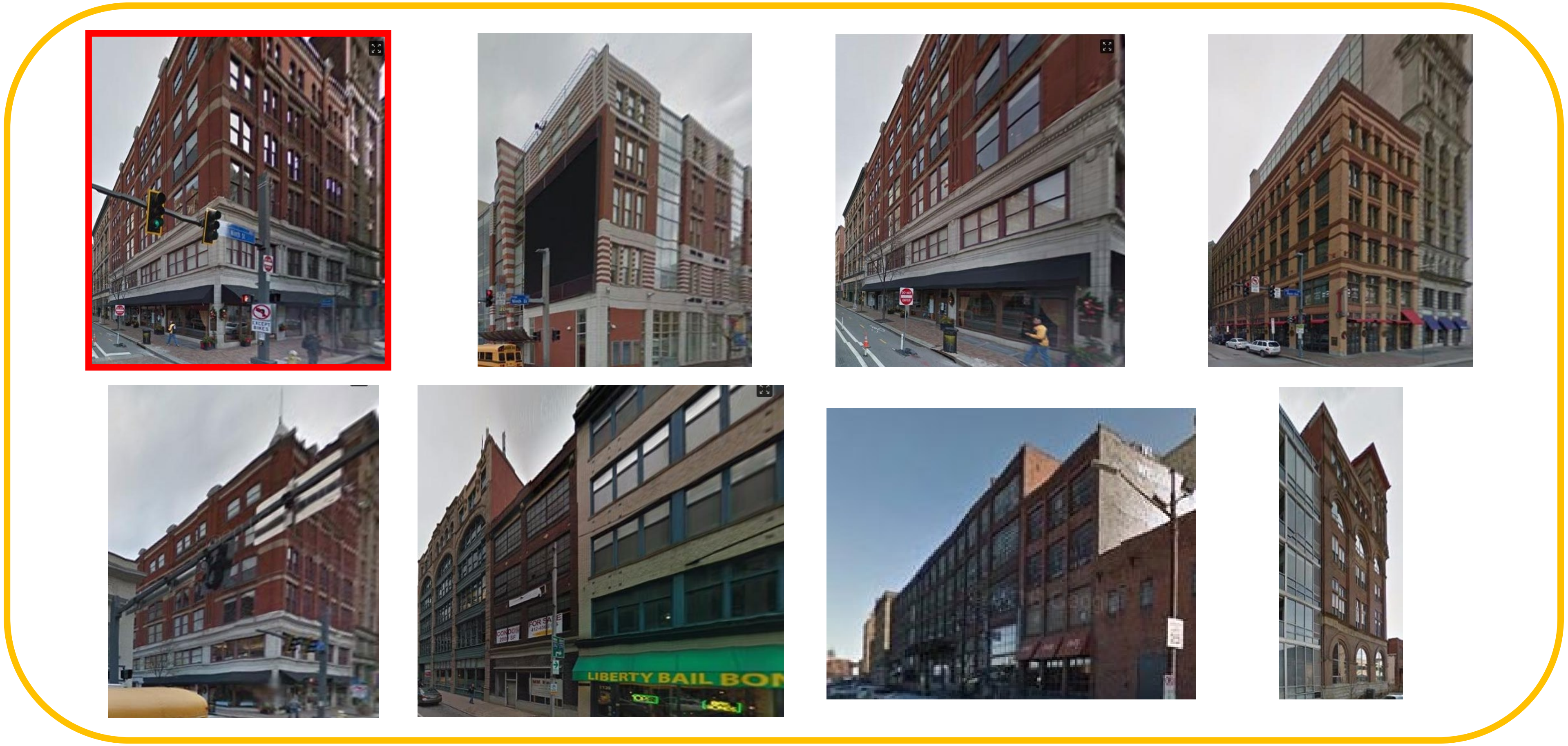


From ~40,000 candidate image patches (Buildings)
(bird’s eye view)

Building matching

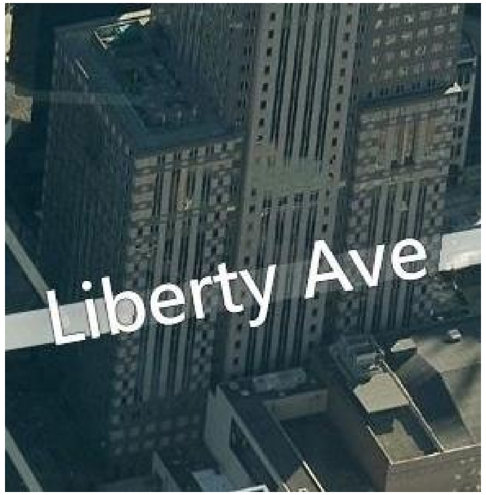
UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Query image (bird's eye view)



From ~10,000 candidate image patches (Buildings) (street view)

Building matching

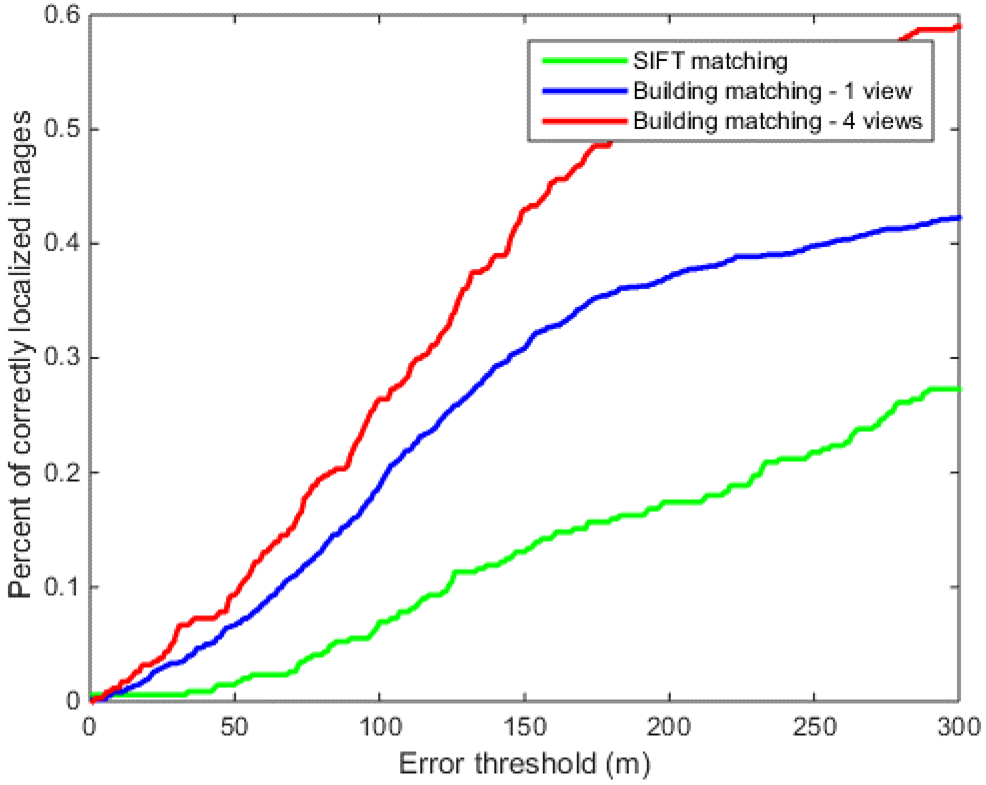


Query image
(bird's eye view)

From ~10,000 candidate image patches (Buildings)
(street view)

Geo-localization

- Query image: street view
- Reference images: bird's eye view
- k: 100



Summary

- Geo-localization Using Cross View Image Matching
  - Detect Buildings
  - Match Buildings
  - Retrieve k-nearest neighbors for each query
  - Dominant Set Selection

Cross-View Image Matching for Geo-localization in Urban Environments

Yicong Tian, Chen Chen, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida

CVPR-2017

Retrieval Features

- Local Features (SIFT)
- Building Features
- **Global Image Features**

Bridging the Domain Gap for Ground-to-Aerial Image Matching

Krishna Regmi

Krishna Regmi & Mubarak Shah
University of Central Florida

https://www.youtube.com/watch?v=gmAhQXCYCEQ&list=PLd3hlSJsX_IkSnnrMtzsMHl1q6vimipvp&index=1

Introduction

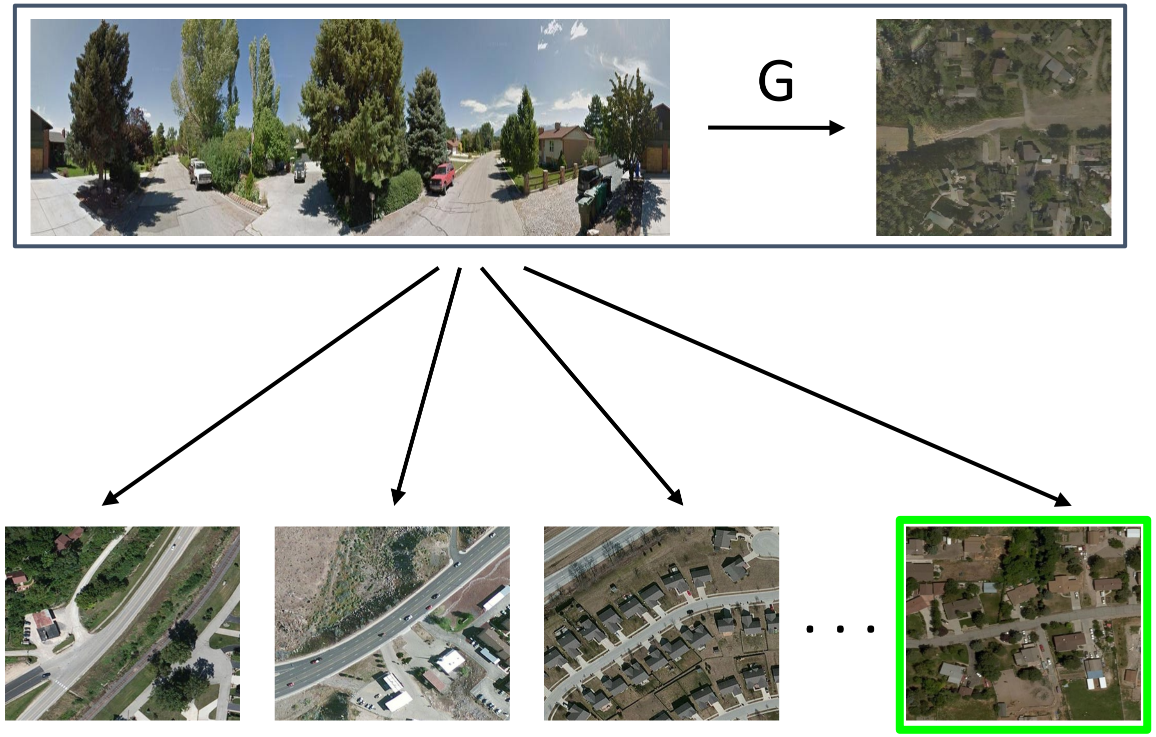
Cross-view image matching

Drastic viewpoint differences

Bridge the domain gap - use GANs
(synthesize target-view images)

Joint Feature Learning and Feature fusion

Multi-scale feature aggregation

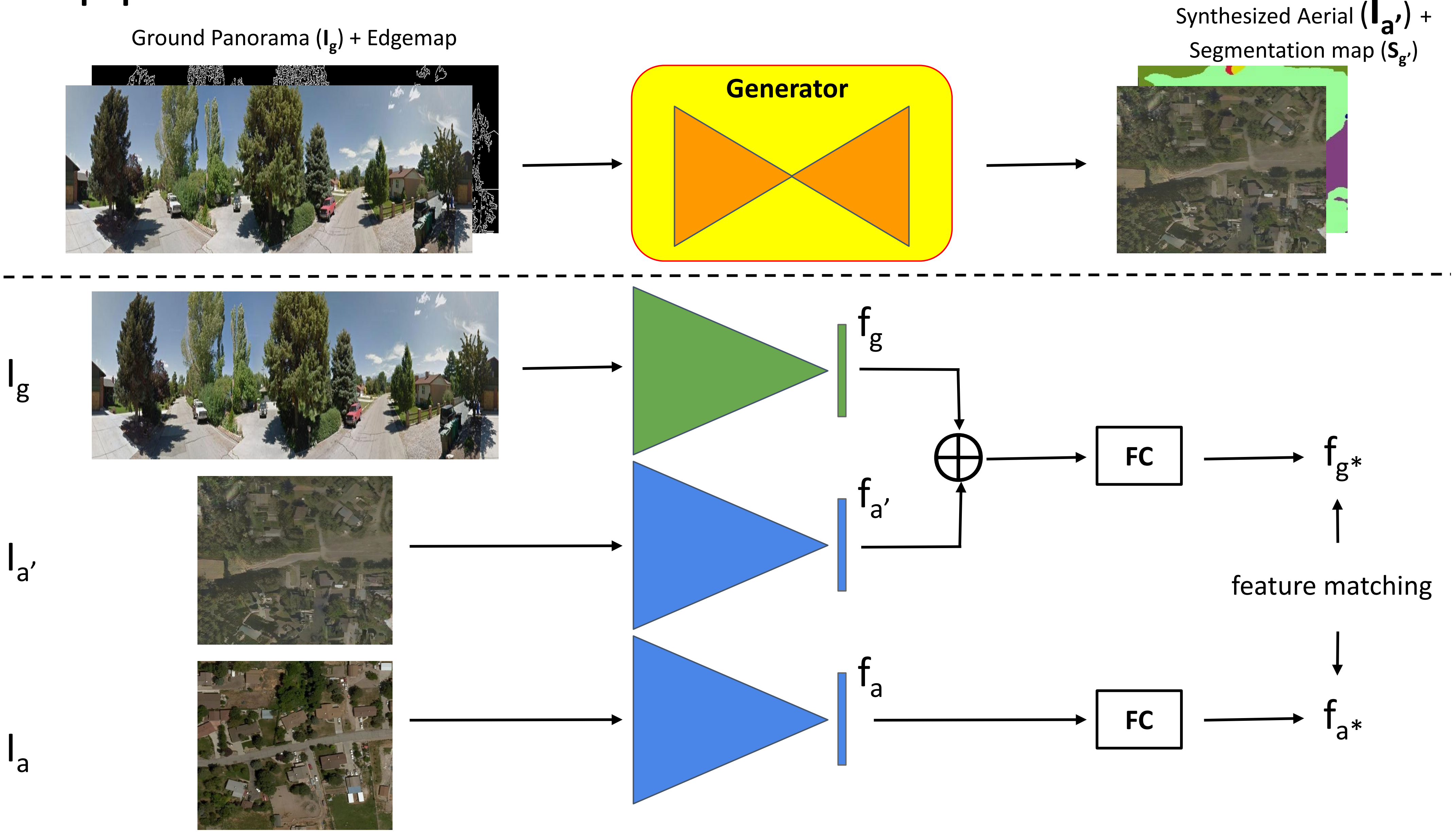


Our Approach

Cross-view image synthesis followed by

Joint Feature Learning and Feature Fusion

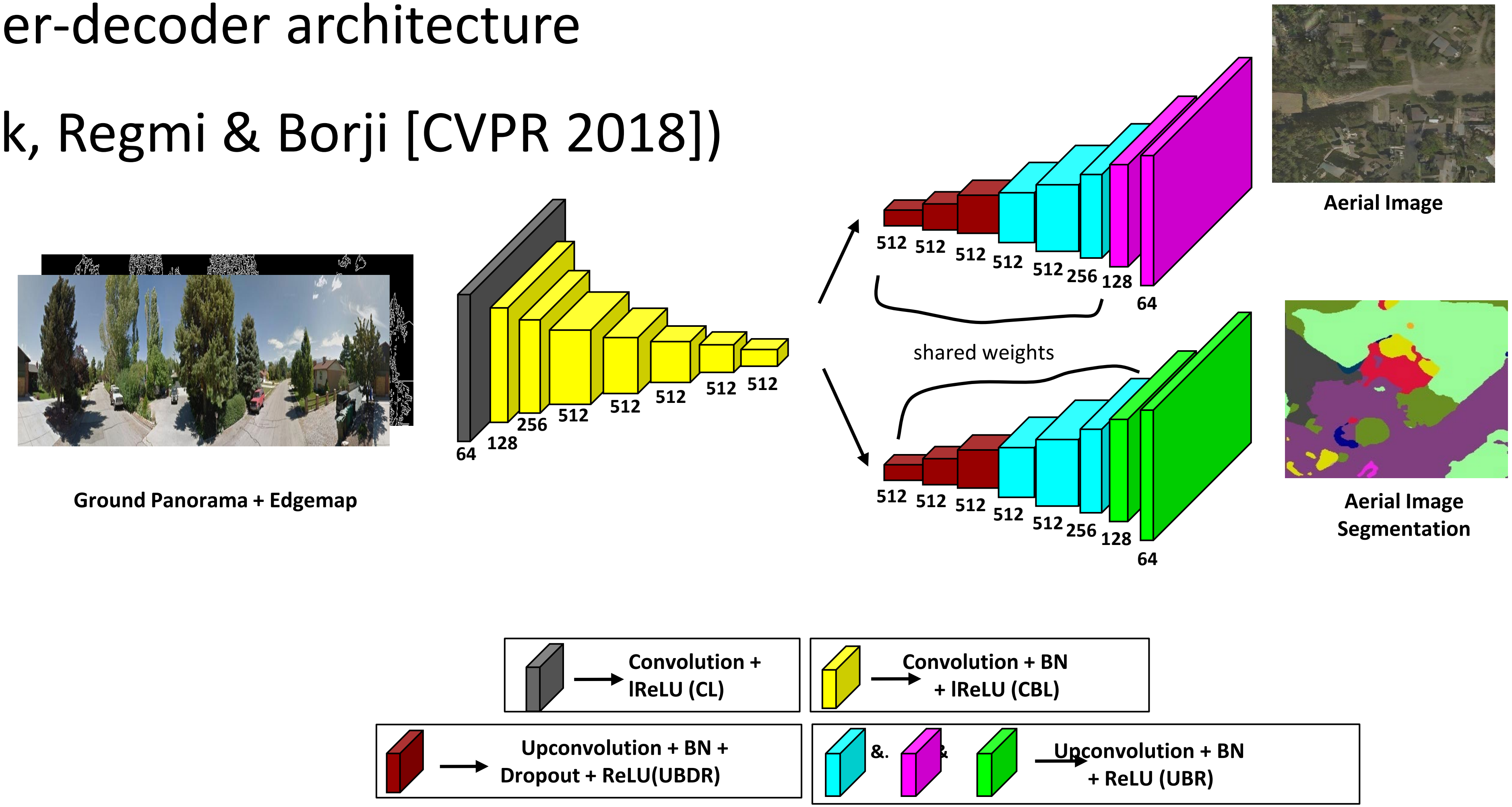
Our Approach



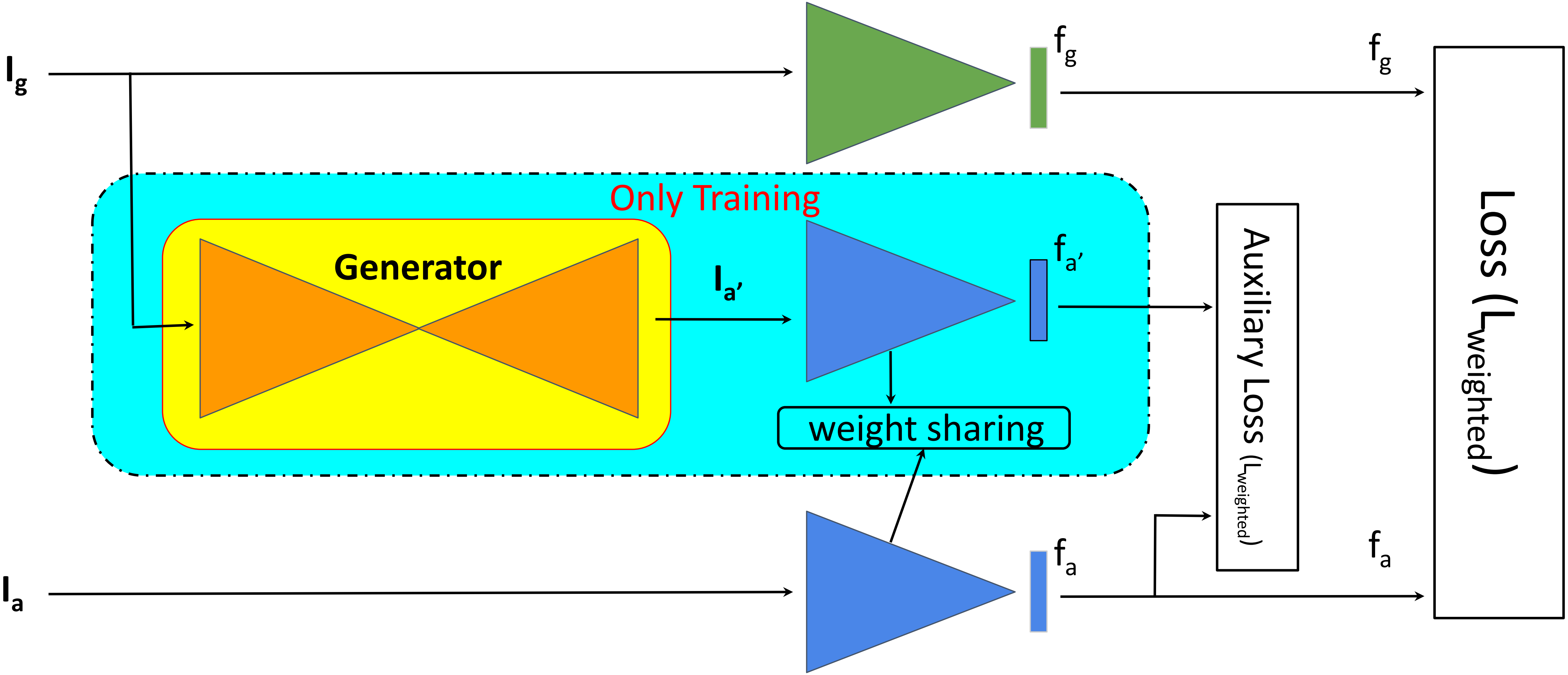
Generator Architecture

Encoder-decoder architecture

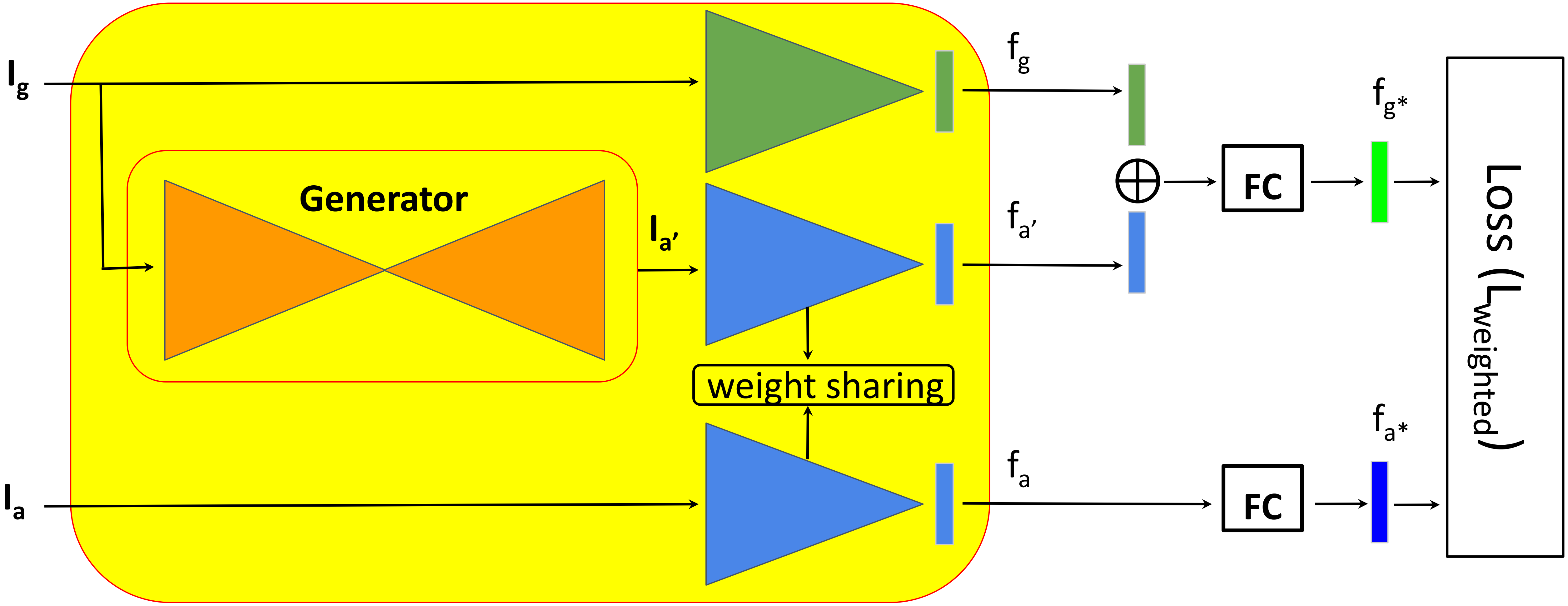
(X-Fork, Regmi & Borji [CVPR 2018])



Joint Feature Learning



Feature Fusion



Loss Functions

Triplet Loss: $$L_{triplet} = max(0, m + d_p - d_n)$$

$$= max(0, m + d) \quad \text{where, } d = d_p - d_n$$

Soft-Margin Triplet Loss: $$L_{soft} = ln(1 + e^{d})$$

Weighted Soft-Margin Triplet Loss: $$L_{weighted} = ln(1 + e^{\alpha d})$$

Loss for Joint Feature Learning:

$$L_{joint} = \lambda_1 L_{weighted}(I_g, I_a) + \lambda_2 L_{weighted}(I_{a'}, I_a)$$

Datasets

Satellite - ground panorama pairs

CVUSA Dataset:

Train/Test: 35,532/8,884 pairs

Covers rural areas

UCF OP Dataset: (Orlando Pittsburgh)

Train/Test: 1,910/722 pairs

Newly collected

Covers urban areas

GPS info available

Results: CVUSA Dataset

Quantitative Results (CVUSA Dataset)

Method	Top-1	Top-10	Top-1%
Two-stream baseline ($I_{a'}$, I_a)	10.23%	35.10%	72.58%
Two-stream baseline (I_g, I_a)	18.45%	48.98%	82.94%
Joint Feat. Learning ($I_{a'}$, I_a)	14.31%	48.75%	86.47%
Joint Feat. Learning (I_g, I_a)	29.75%	66.34%	92.09%
Feature Fusion	**48.75%**	**81.27%**	**95.98%**
Workman et al. [41]	-	-	34.3%
Zhai et al. [46]	-	-	43.2%
Vo and Hays [39]	-	-	63.7%
CVM-Net-I [18]	22.53%	63.28%	91.4%
CVM-Net-II [18]	11.18%	43.51%	87.2%

Recall Accuracy (CVUSA Dataset)

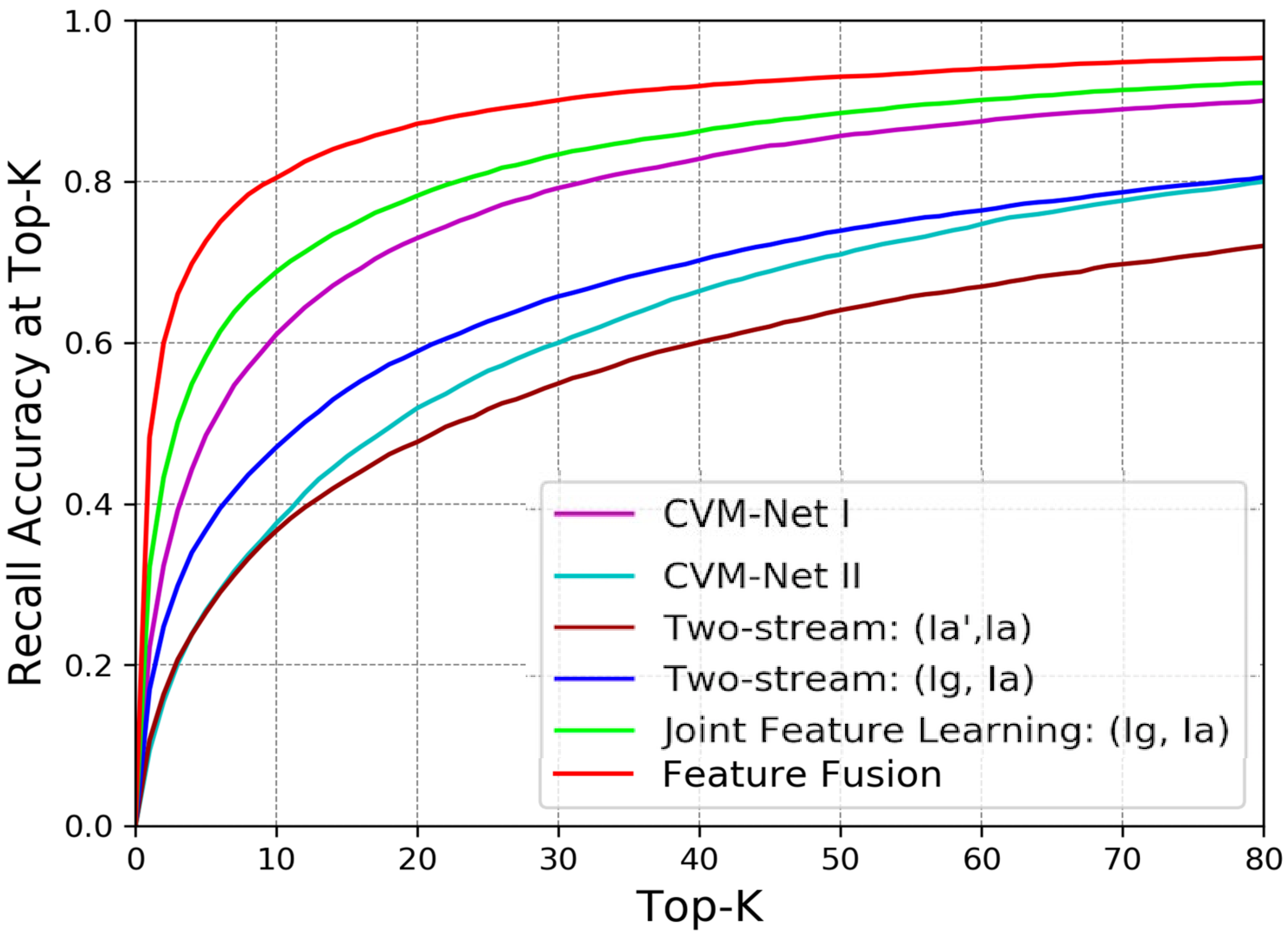
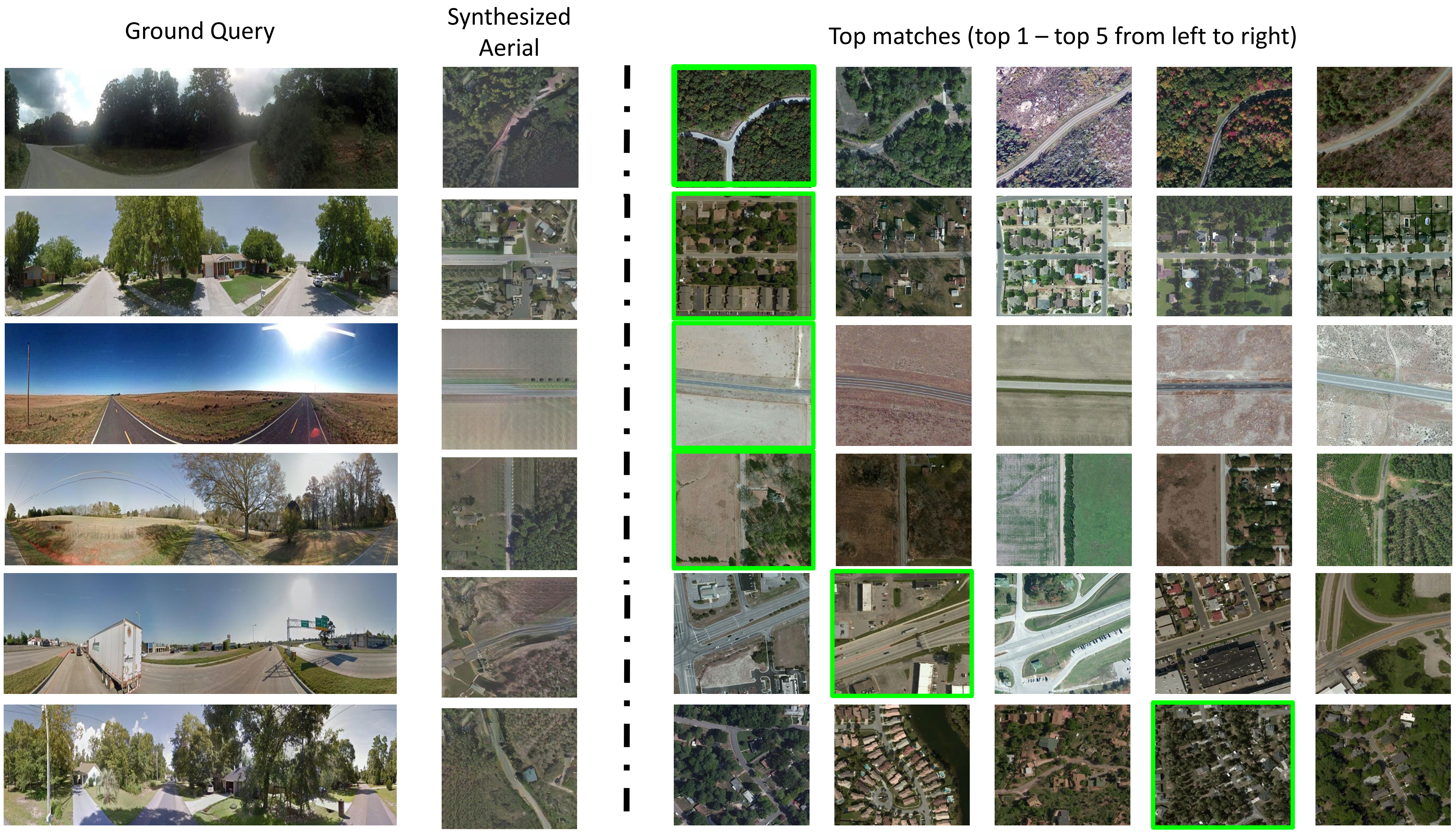


Image Retrieval (CVUSA Dataset)



Failure Cases

Ground Query	Synthesized Aerial	Top-1 match	Correct match (Ground Truth)	Retrieved position
				13
				37
				532
				1700

Results: OP Dataset

Retrieval Performance (OP Dataset)

Two-stream (I_g, I_a)	Joint Feat. Learning	Feature Fusion
30.61%	38.36%	45.57%

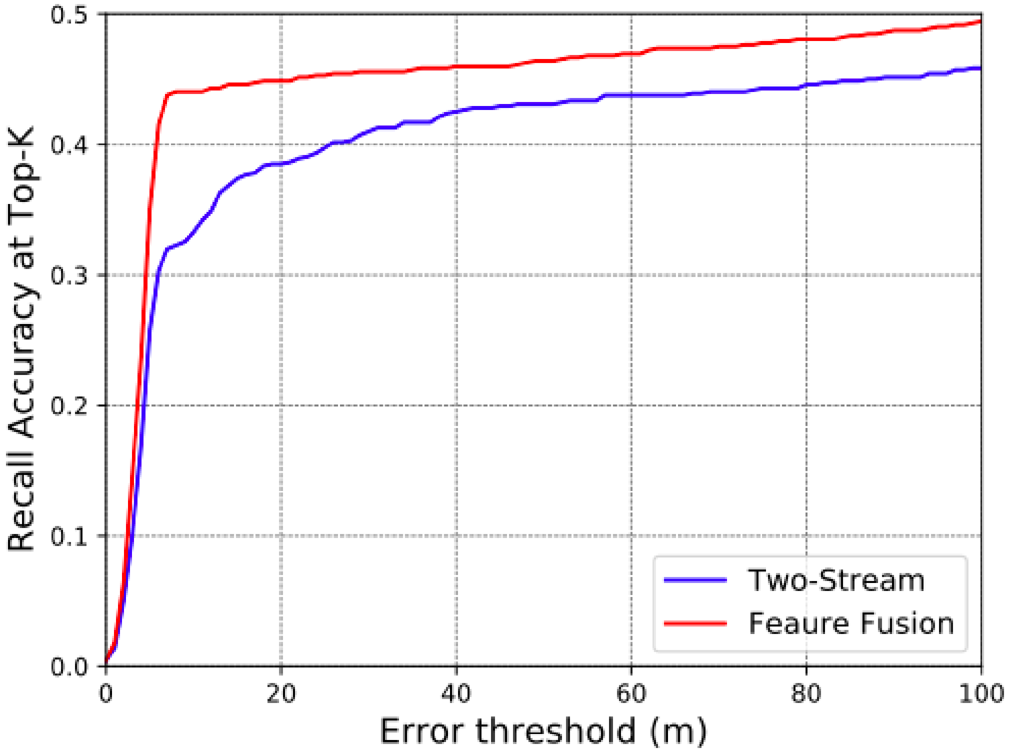
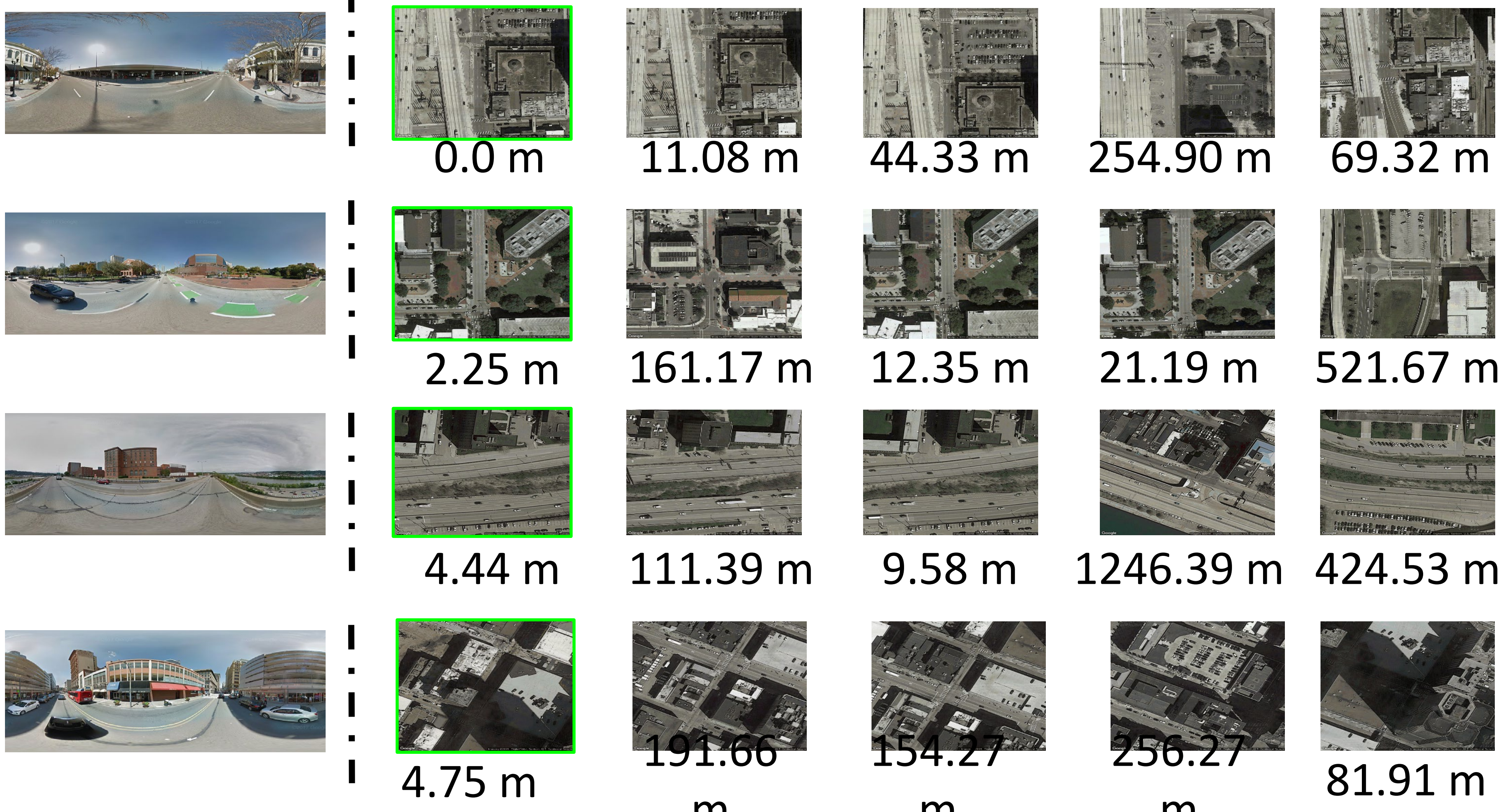


Image Retrieval (OP Dataset)

Ground Query | Top matches (top 1 – top 5 from left to right)



Aerial-to-Ground (A2G) Image Matching

Reverse problem

Synthesize the ground panorama from aerial image

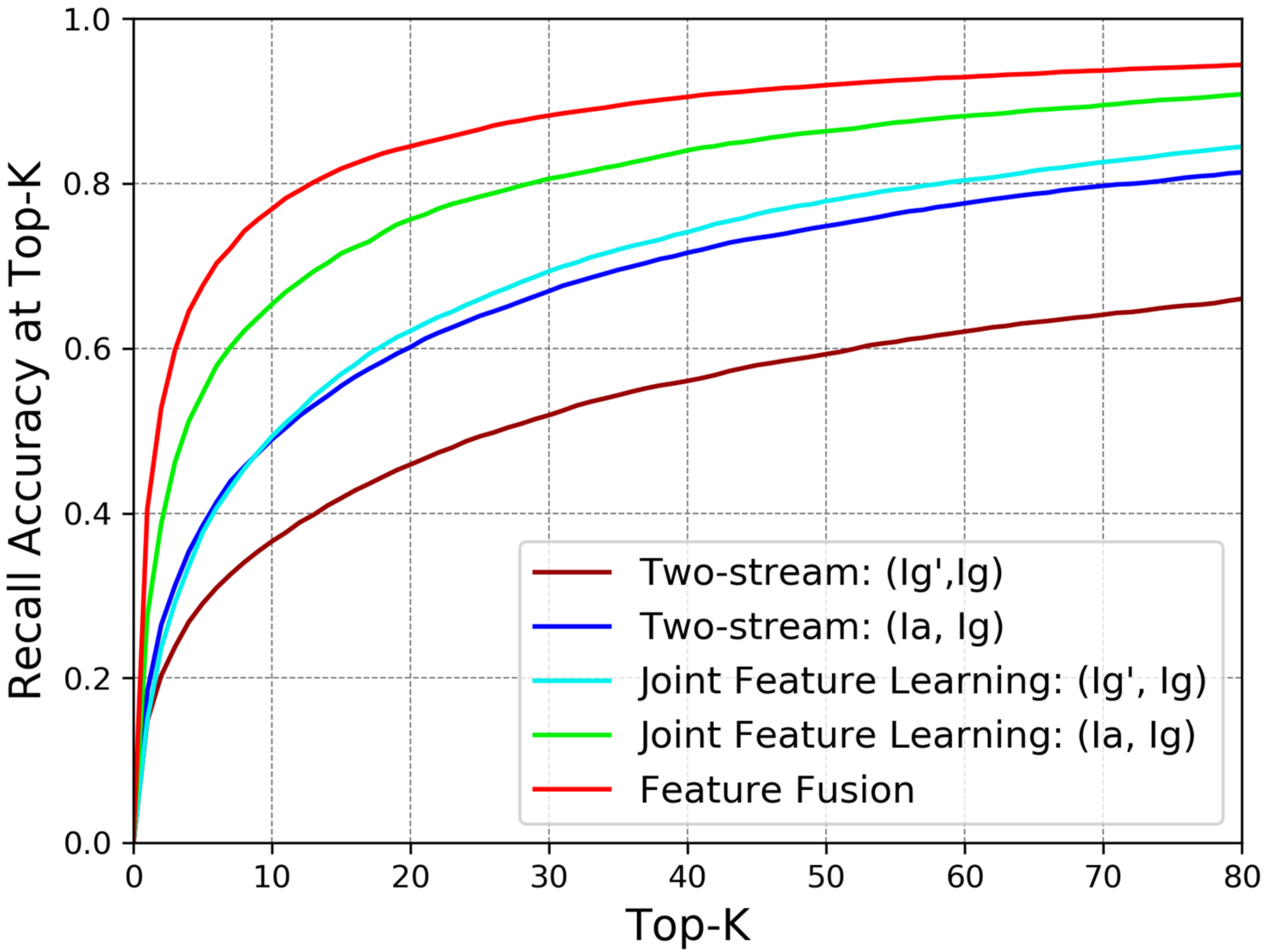
Quantitative Results - A2G (CVUSA Dataset)

Method	Top-1	Top-10	Top-1%
Two-stream baseline ($I_{g'}$, I_g)	15.04%	37.31%	67.99%
Two-stream baseline (I_a, I_g)	16.99%	47.06%	82.11%
Joint Feat. Learning ($I_{g'}$, I_g)	16.46%	50.26%	86.26%
Joint Feat. Learning (I_a, I_g)	27.39%	65.29%	91.46%
Feature Fusion	**44.99%**	**79.37%**	**95.66%**

Image Retrieval - A2G (CVUSA Dataset)

Aerial Query | Synthesized Ground Panorama | Top matches (top 1 – top 5 from left to right)

Recall Accuracy - A2G (CVUSA Dataset)



Summary

- Novel and practical approach to cross-view image matching
- Domain gap is bridged by synthesized images
- Significant improvement on Top-1 and Top-10 accuracies over SOTA on CVUSA.
- This approach can be used for other view transformation tasks where the transformations can be in horizontal or vertical directions.

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

ICCV 2019
Seoul, Korea

Bridging the Domain Gap for Ground-to-Aerial Image Matching

Krishna Regmi

Krishna Regmi & Mubarak Shah
University of Central Florida

Summary

- Pixel-Wise Geo-localization
  - Geodetic Alignment of Aerial Video Frames
- Image-Based Geo-Localization
  - Same View (Street-View to Street-View)
    - Generalized Maximum Clique (PAMI, 2014)
    - Constraint Dominant Sets (PAMI, 2017)
  - Cross-View Geo-Localization
    - Bird’s Eye-View to Street View (CVPR, 2017)
    - Aerial to Ground View (ICCV, 2019)

Thank You