Deep Geolocalization and Siamese Nets

Computer Vision

James Hays

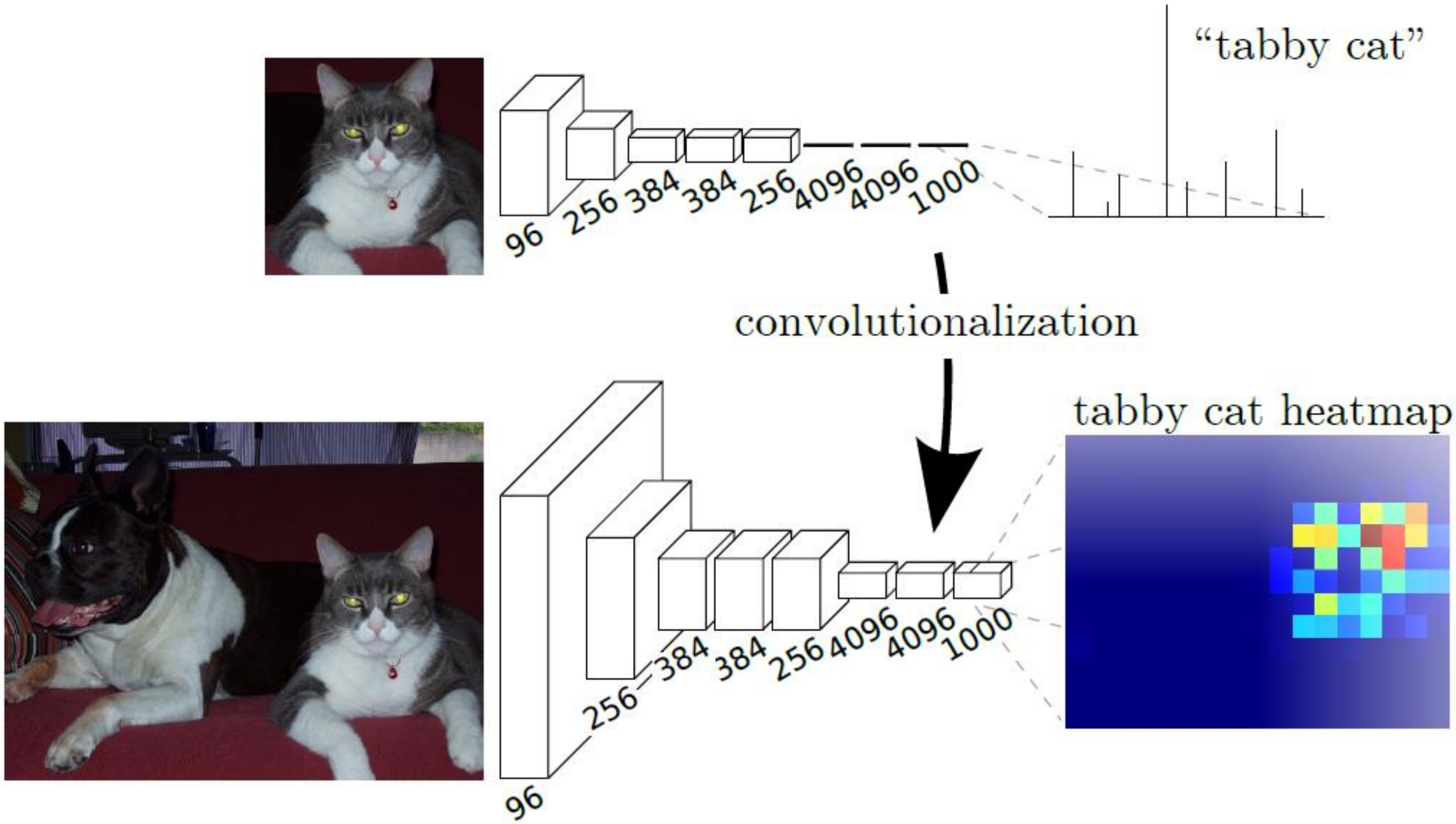
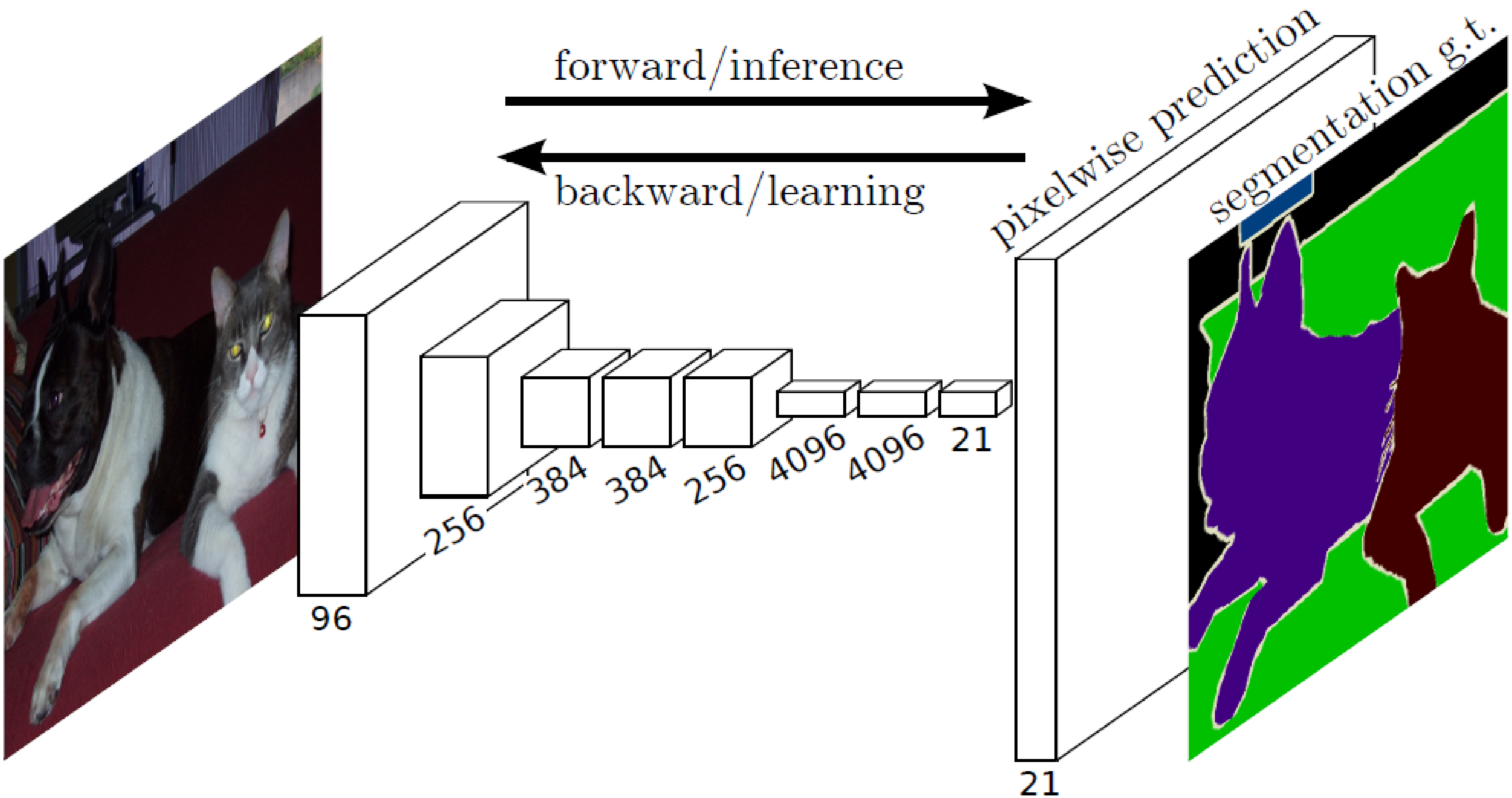


Figure 2. Transforming fully connected layers into convolution layers enables a classification net to output a heatmap. Adding layers and a spatial loss (as in Figure 1) produces an efficient machine for end-to-end dense learning.

Long, Shelhamer, and Darrell 2014



Long, Shelhamer, and Darrell 2014

PlaNet - Photo Geolocation with Convolutional Neural Networks

Tobias Weyand, Ilya Kostrikov, James Philbin

ECCV 2016

Discretization of Globe

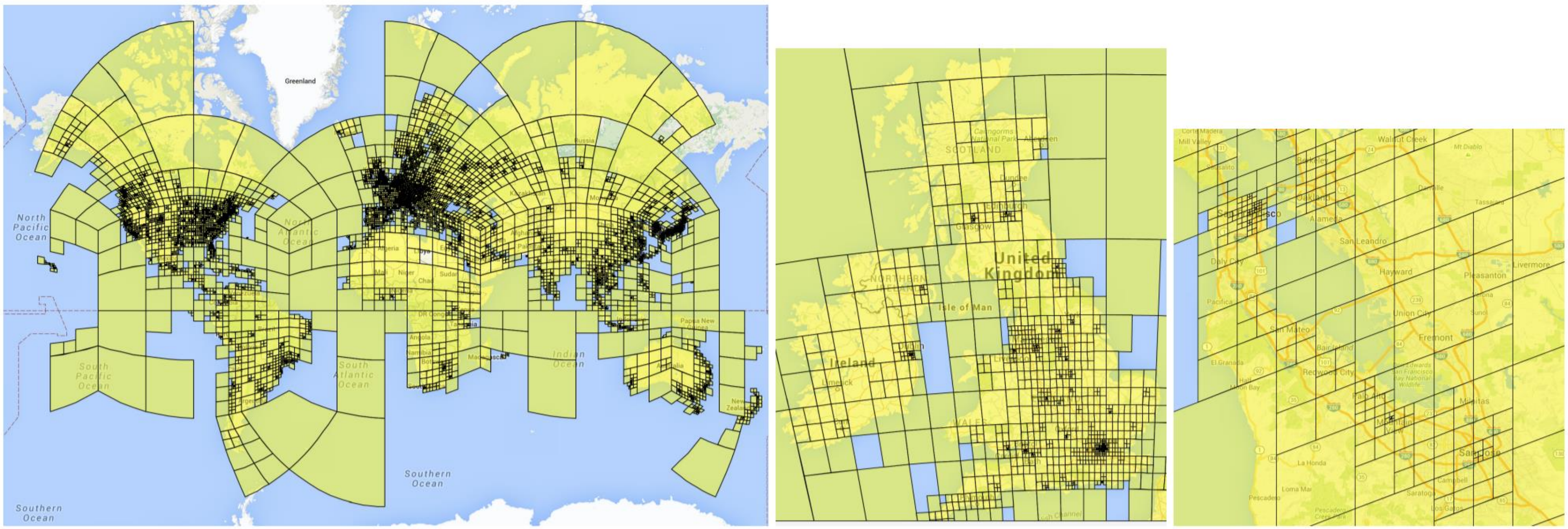


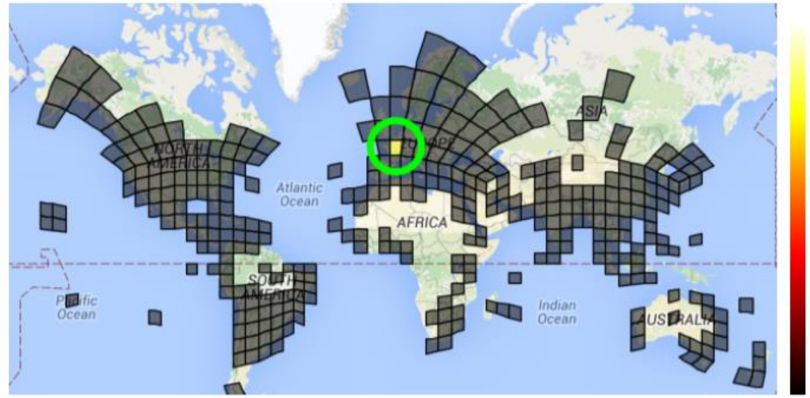
Figure 2. Left: Adaptive partitioning of the world into 26,263 S2 cells. Right: Detail views of Great Britain and Ireland and the San

Network and Training

- Network Architecture: Inception with 97M parameters
- 26,263 “categories”

- 126 Million Web photos
- 2.5 months of training on 200 CPU cores

Photo CC-BY-NC by stevekc



(a)

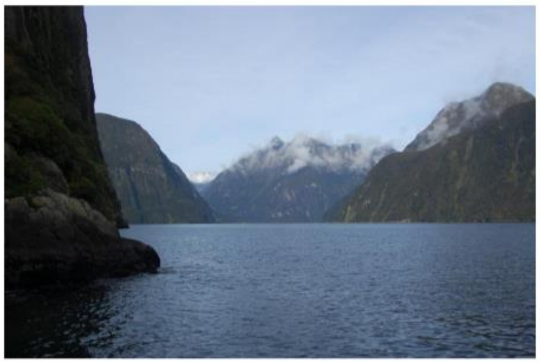
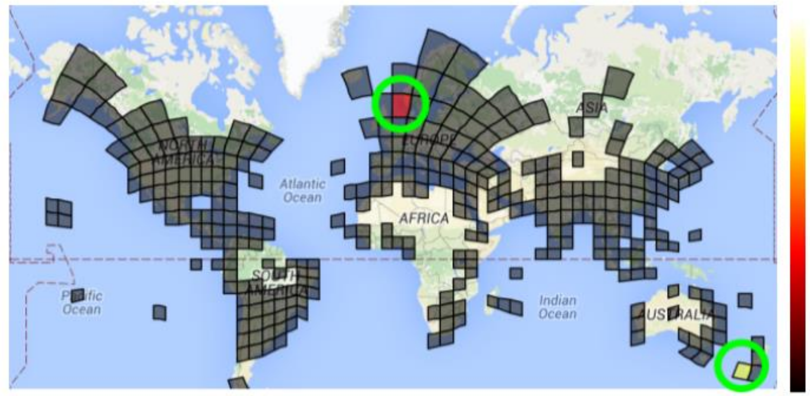


Photo CC-BY-NC by edwin.11



(b)

Photo CC-BY-NC by jonathanfh

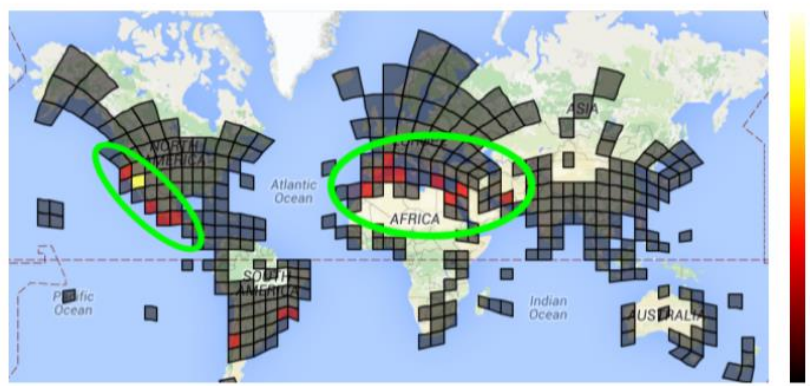


Photo by jamie.lovelock / CC BY NC Photo by MongoosePhotography / CC BY NC

Photo by Mister-E / CC BY NC Photo by dalangalma / CC BY NC Photo by siamjack / CC BY NC

Namibia / Botswana

Photo by ryan + sarah / CC BY NC Photo by stuartlchambers / CC BY NC Photo by samgrover / CC BY NC

Photo by steuben / CC BY NC Photo by steve-stevens / CC BY NC

Kauai, Hawaii

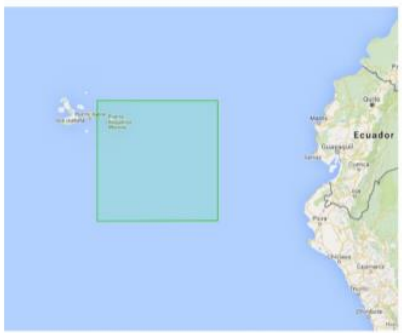


Photo by p.j.k. / CC BY NC Photo by victor408 / CC BY NC Photo by Domen Jakus / CC BY NC

Photo by cvanholder / CC BY NC Photo by rwoan / CC BY NC

Galapagos Islands

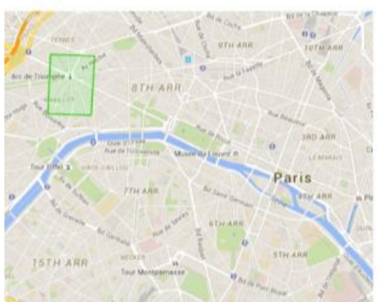


Photo by feliven / CC BY NC Photo by fred_v / CC BY NC Photo by Turansa Tours / CC BY NC

Photo by JA_FS / CC BY NC Photo by CedEm photographies / CC BY NC

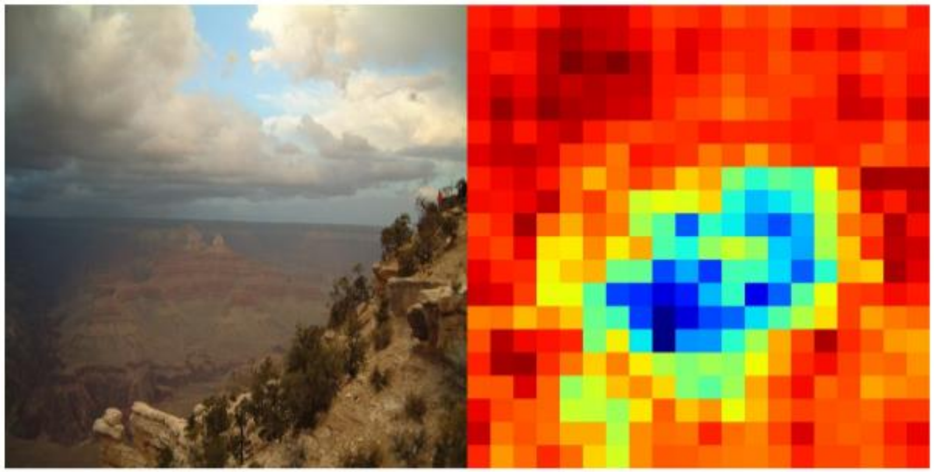
Paris

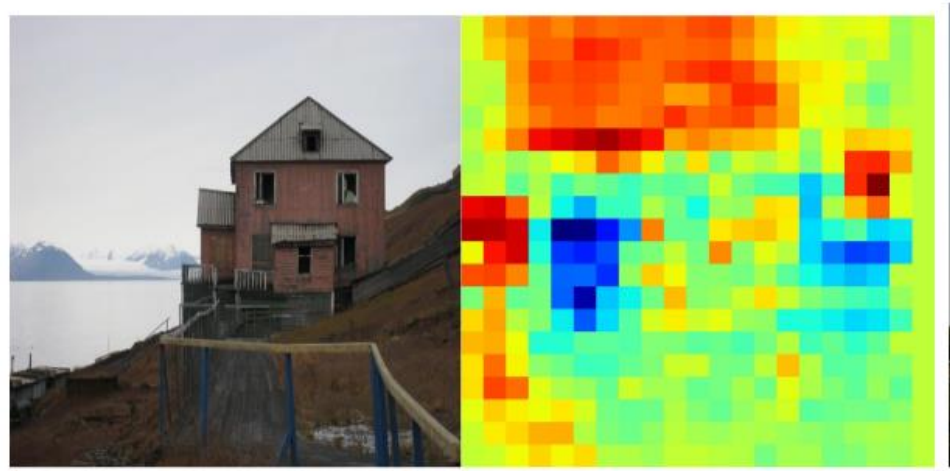
PlaNet vs im2gps (2008, 2009)

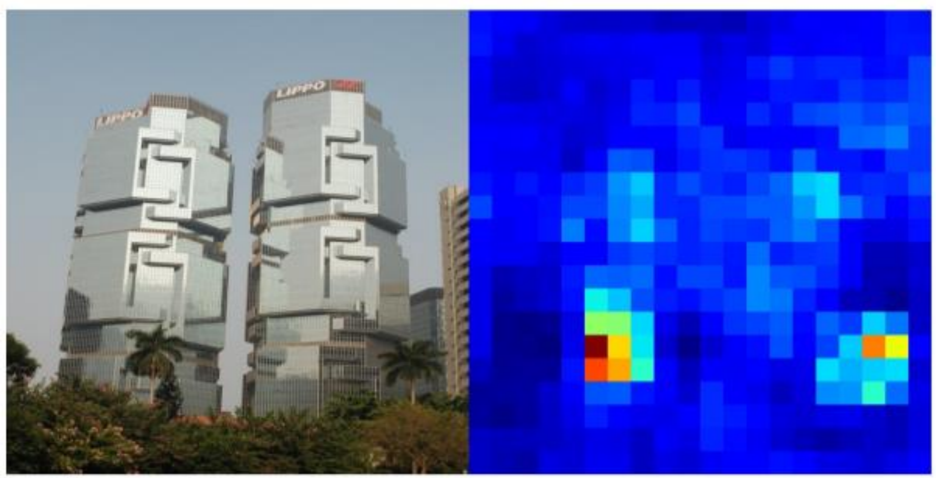
Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS (orig) [17]		12.0%	15.0%	23.0%	47.0%
Im2GPS (new) [18]	2.5%	21.9%	32.1%	35.4%	51.9%
PlaNet	**8.4%**	**24.5%**	**37.6%**	**53.6%**	**71.3%**

Method	Manmade Landmark	Natural Landmark	City Scene	Natural Scene	Animal
Im2GPS (new)	61.1	37.4	3375.3	5701.3	6528.0
PlaNet	74.5	61.0	212.6	1803.3	1400.0

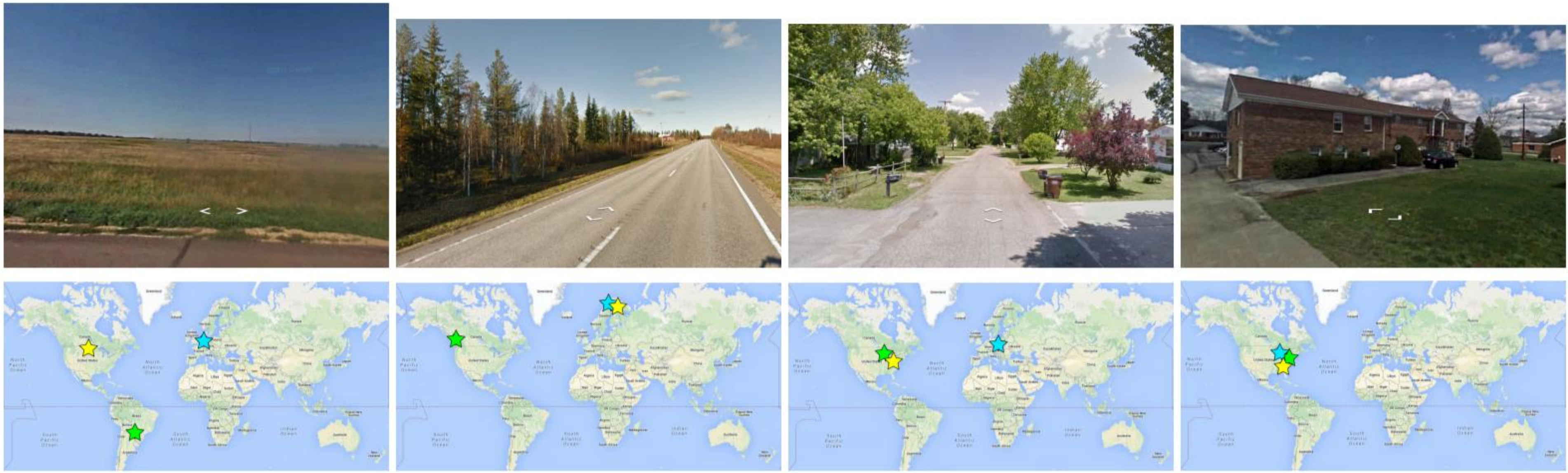
Spatial support for decision



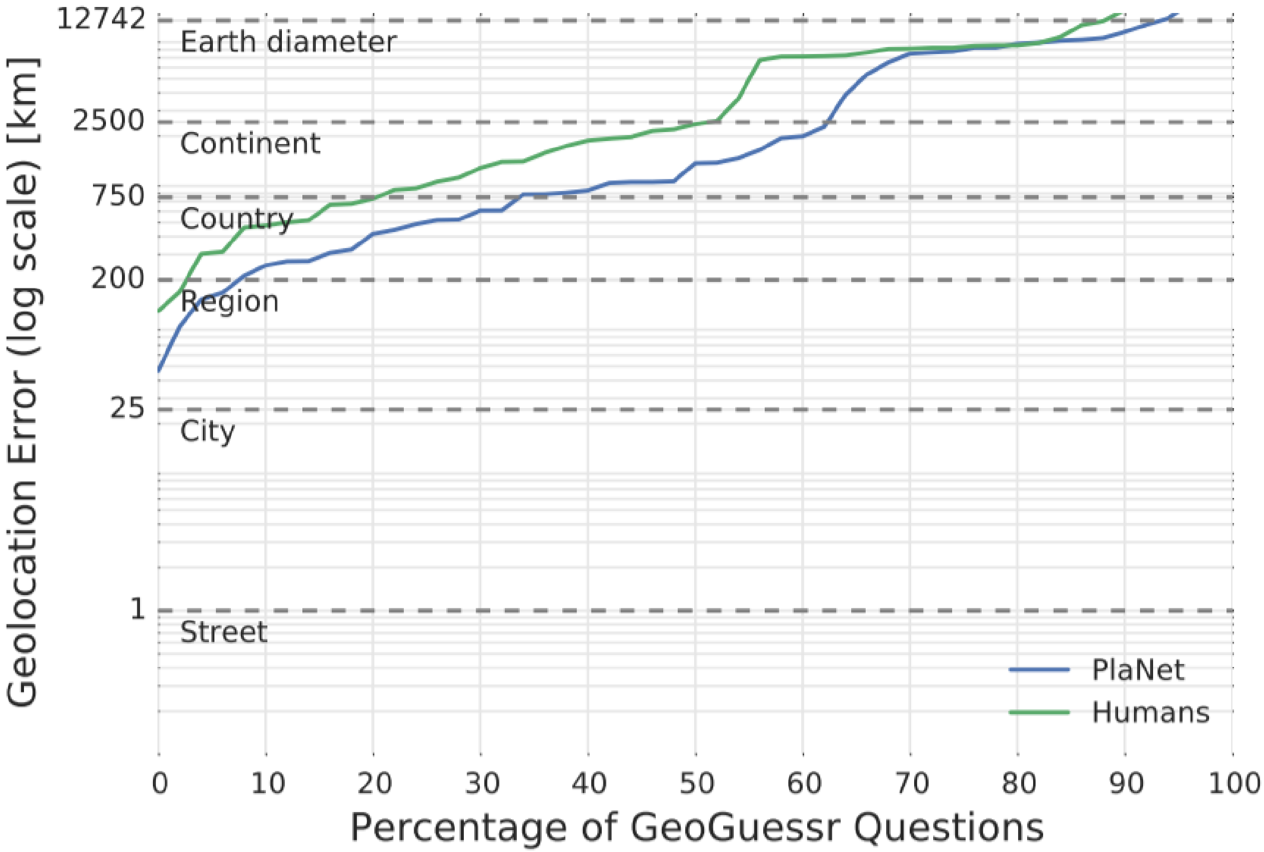




PlaNet vs Humans



PlaNet vs Humans



PlaNet summary

- Very fast Geolocalization method. Geolocalization by categorization.
- Uses far more training data than previous work (im2gps)
- There’s definitely still room for improvement

Learning Deep Representations For Ground-to-Aerial Geolocalization

Tsung-Yi Lin, Yin Cui, Serge Belongie, James Hays

CVPR 2015

View From Your Window Contest

June 9, 2010 – Feb. 4, 2015

Where was the photo taken?

Ans:
Milano, Italy

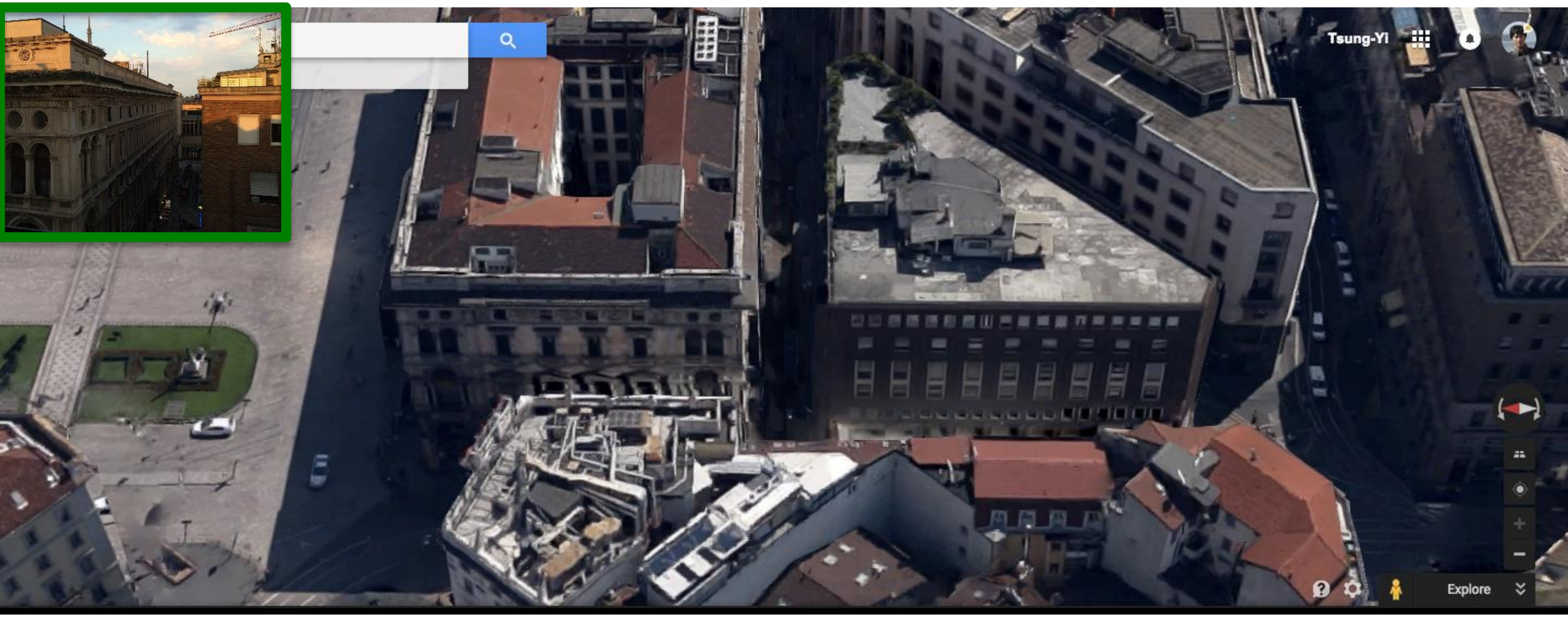
To Geolocalize a Photo

- One can capture every corner on the earth




...

To Geolocalize a Photo




Tsung-Yi
Explore


Tsung-Yi
Explore

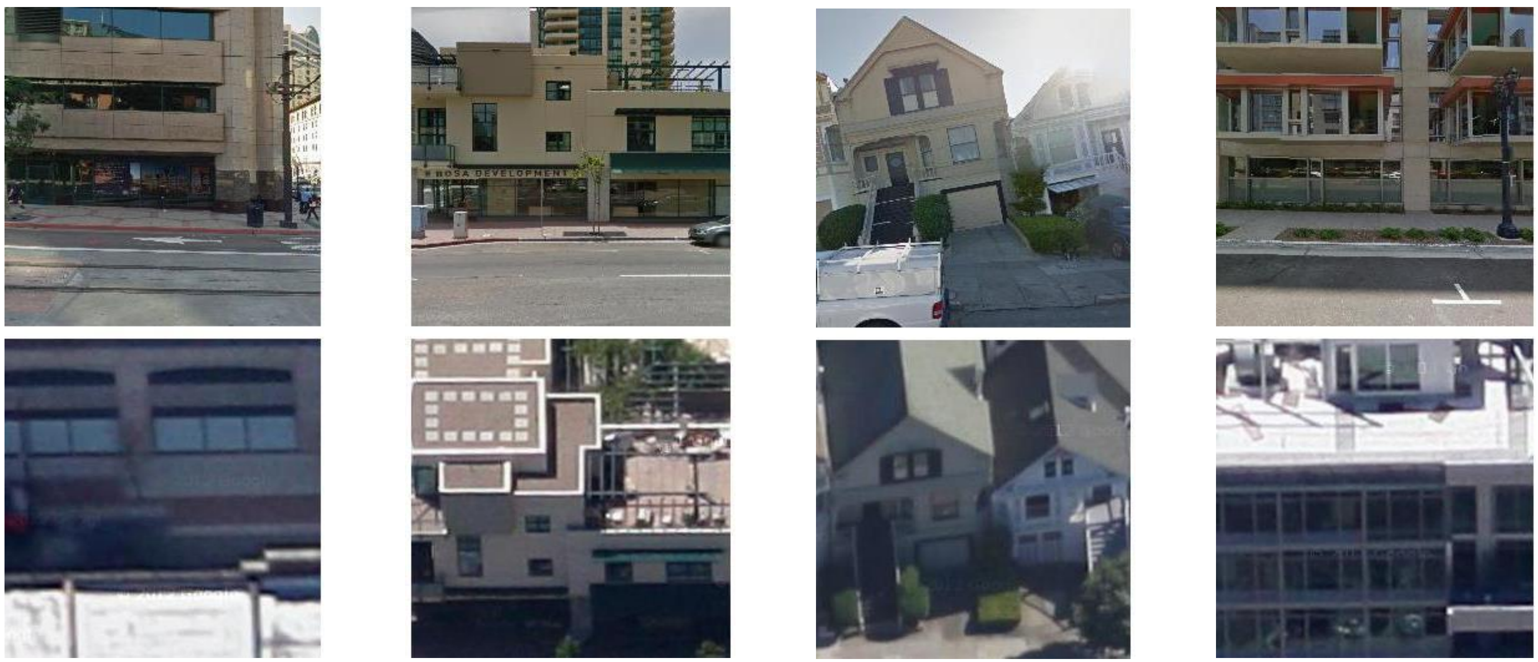
How To Match Ground-to-Aerial?

Shan et al., Accurate Geo-registration by Ground-to-Aerial Image Matching, 3DV’14
Bansal et al., Ultra-wide baseline façade matching for geo-localization, ECCV workshop’12

Are these the same location?

Ground

Aerial



Are these the same location?

Ground

Aerial

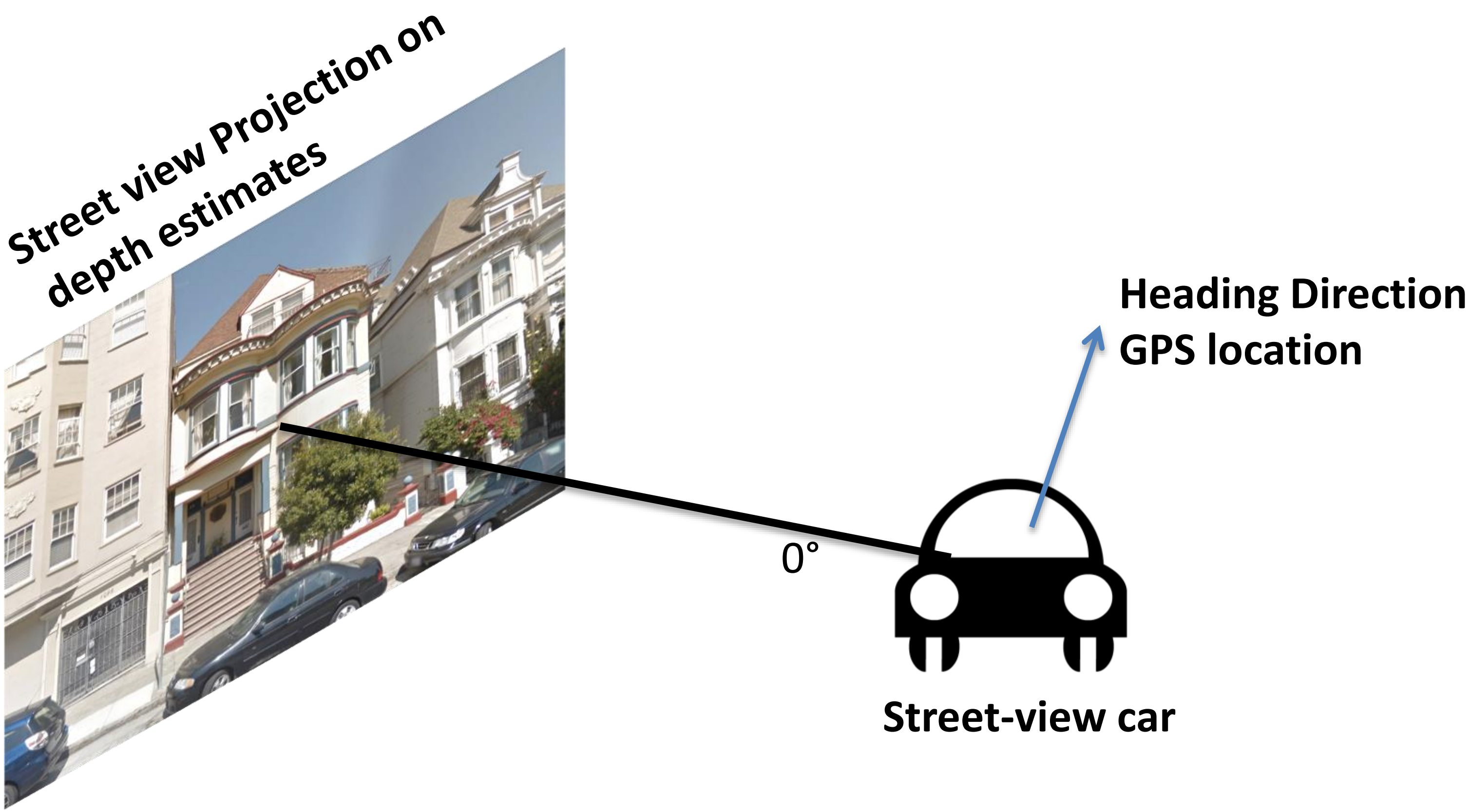
Why Don’t You Just…

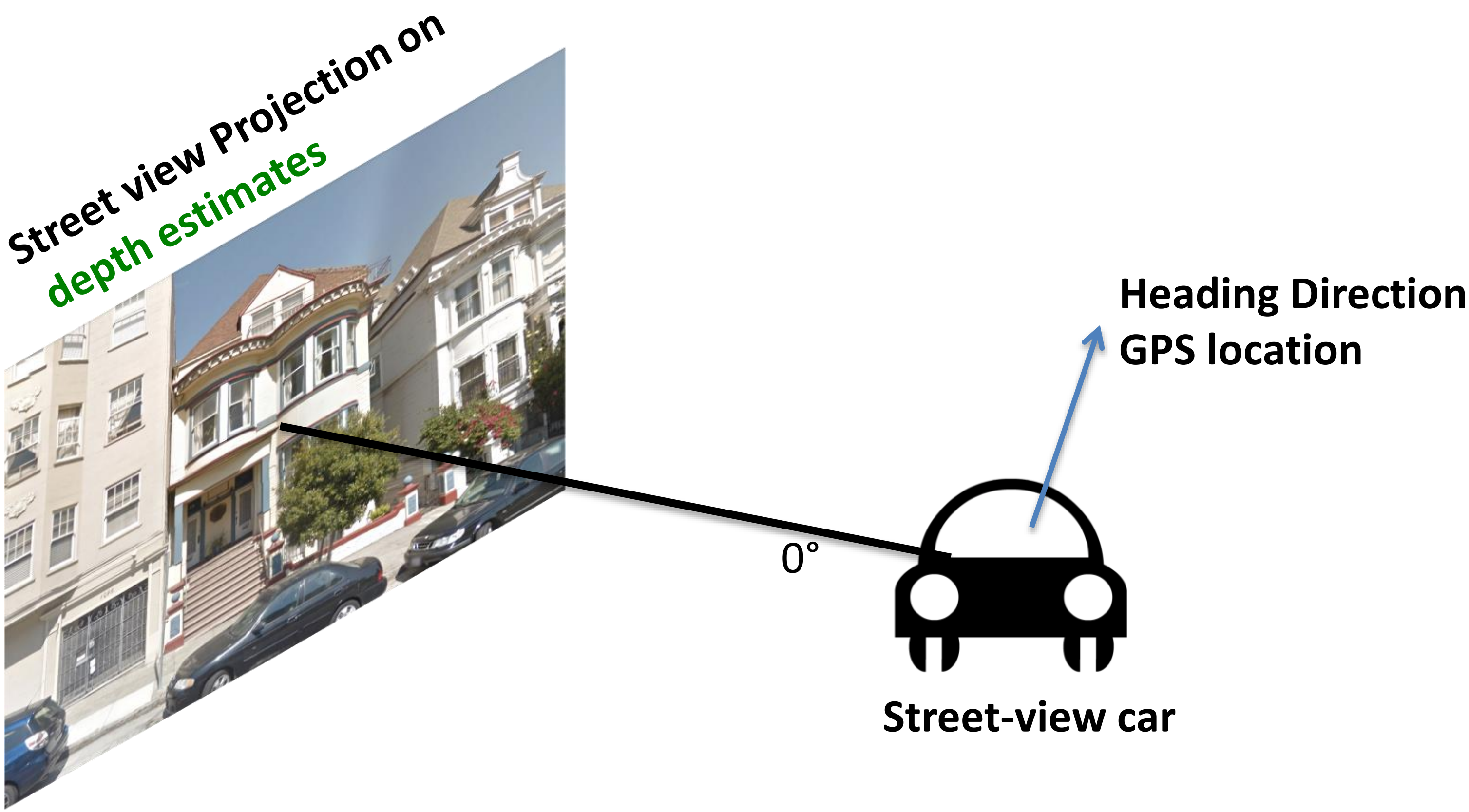
- Sparse Keypoint Matching + RANSAC

Cross-view Pairs

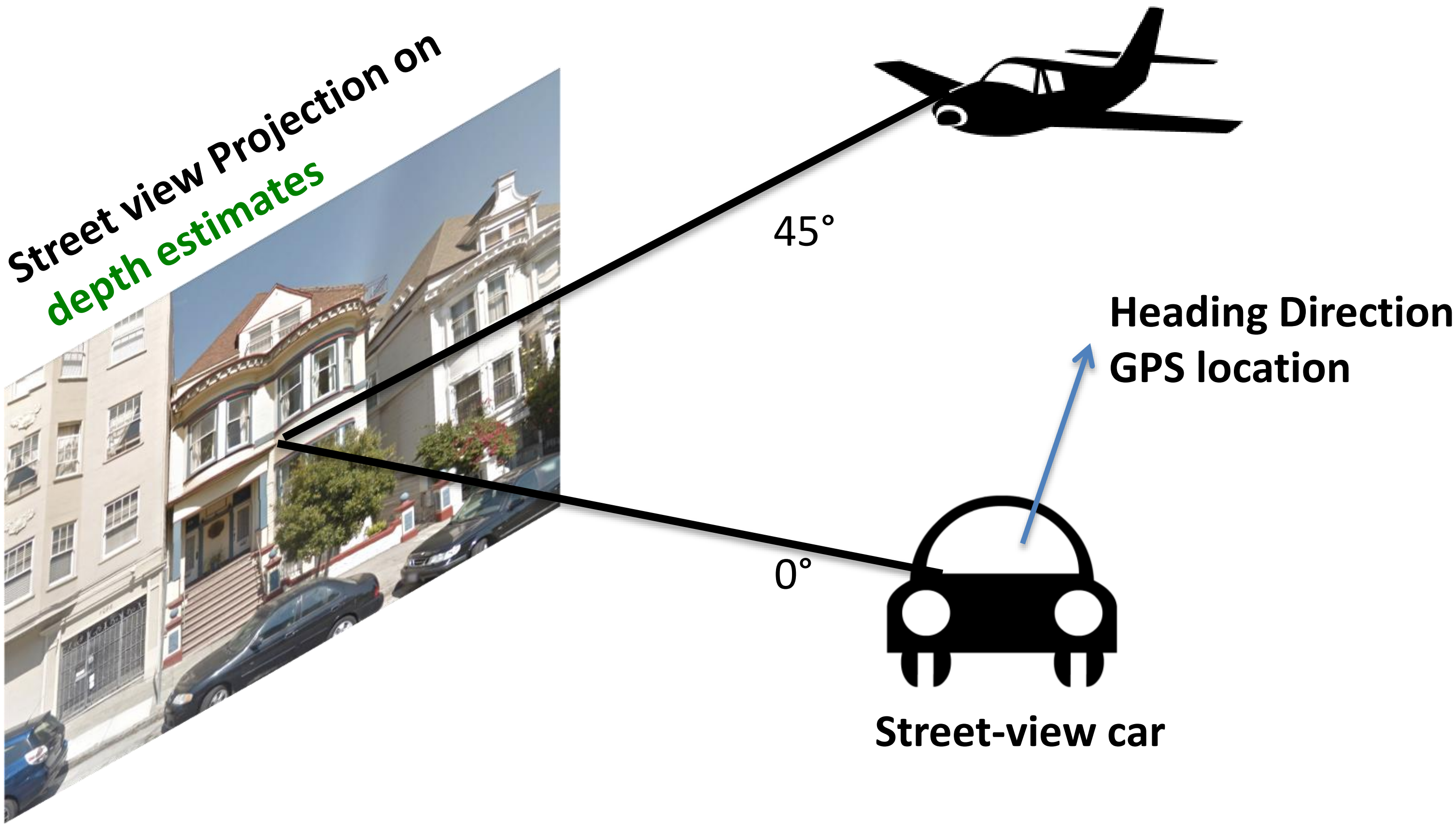
Ground

Aerial


Street view Projection on
depth estimates
Heading Direction
GPS location
0°
Street-view car



Street view Projection on
depth estimates
0°
Heading Direction
GPS location
Street-view car



Street view Projection on
depth estimates
45°
0°
Heading Direction
GPS location
Street-view car

7 Cities: 78k Corresponding Pairs

San Francisco

San Diego

Chicago

Charleston

Tokyo

Rome

Lyon

Place Verification

Same

OR

Different

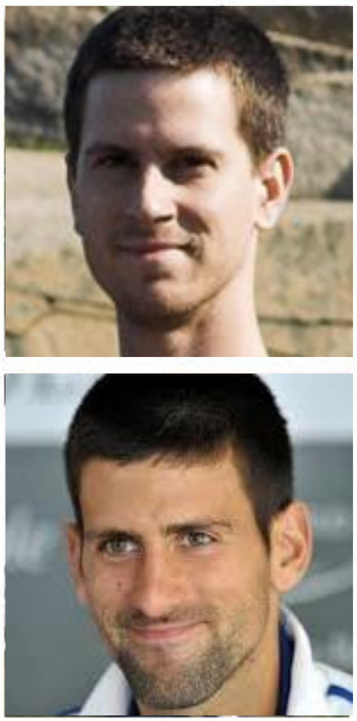
Face Verification

Same

OR

Different

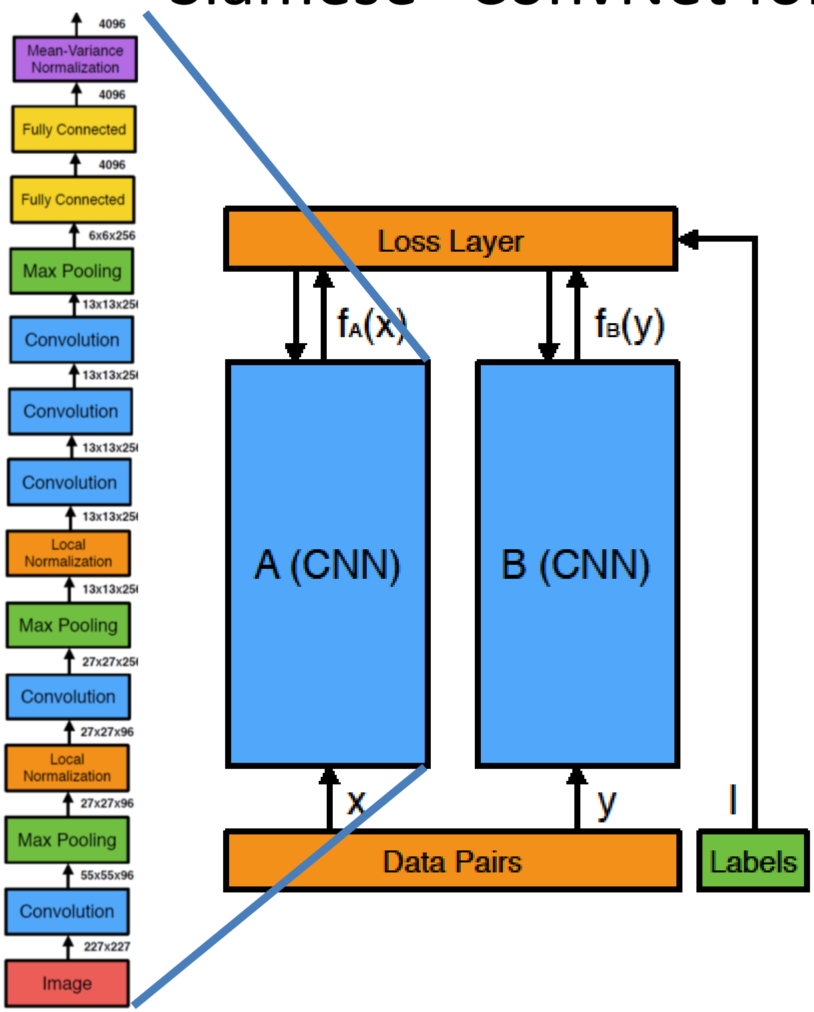
Face Verification



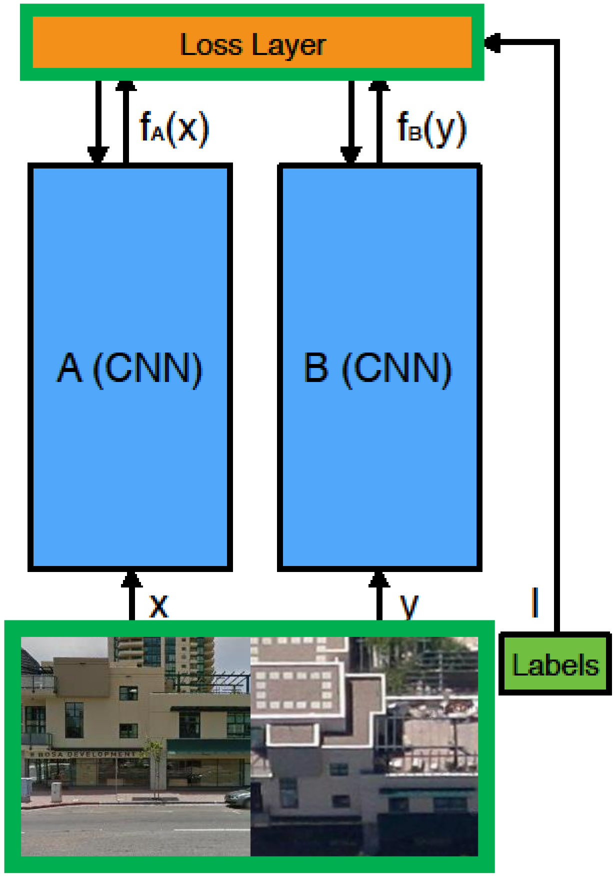
Huang et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments

- Chopra and Hadsell and LeCun, **Learning a Similarity Metric Discriminatively, with Application to Face Verification** (CVPR 2005)
- Taigman, Yang, Ranzato, Wolf, **DeepFace: Closing the Gap to Human-Level Performance in Face Verification** (CVPR 2014)
- Schroff, Kalenichenko, Philbin, **FaceNet: A Unified Embedding for Face Recognition and Clustering** (CVPR 2015)

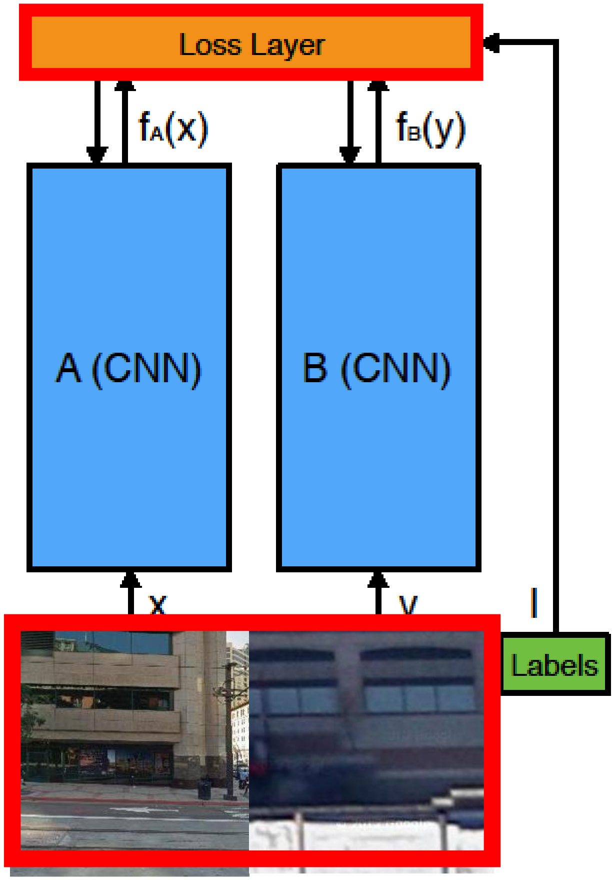
“Siamese” ConvNet for Ground-to-Aerial Matching

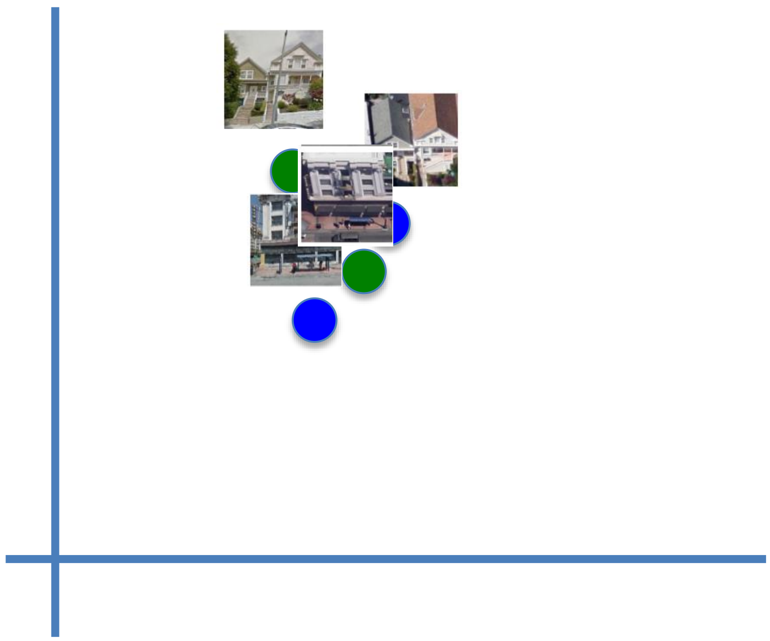


“Siamese” ConvNet for Ground-to-Aerial Matching



“Siamese” ConvNet for Ground-to-Aerial Matching





Contrastive Loss

Loss Function:

- **For similar pairs:**

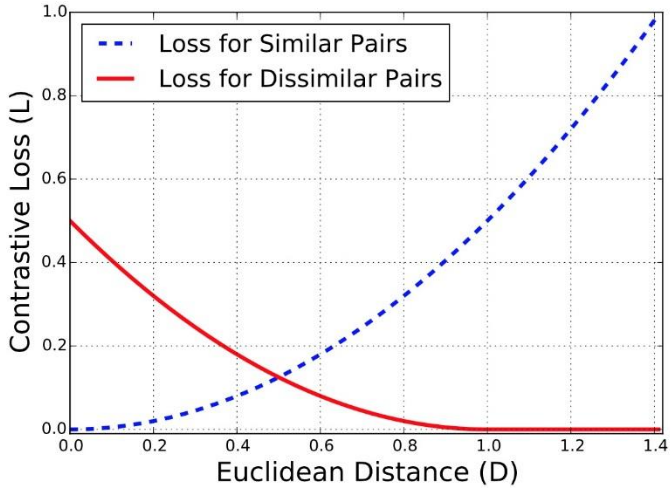
||f()-f() ||^2

- **For dissimilar pairs**

max(0, m- ||f()-f() ||)^2

red: similar pairs

blue: dissimilar pairs

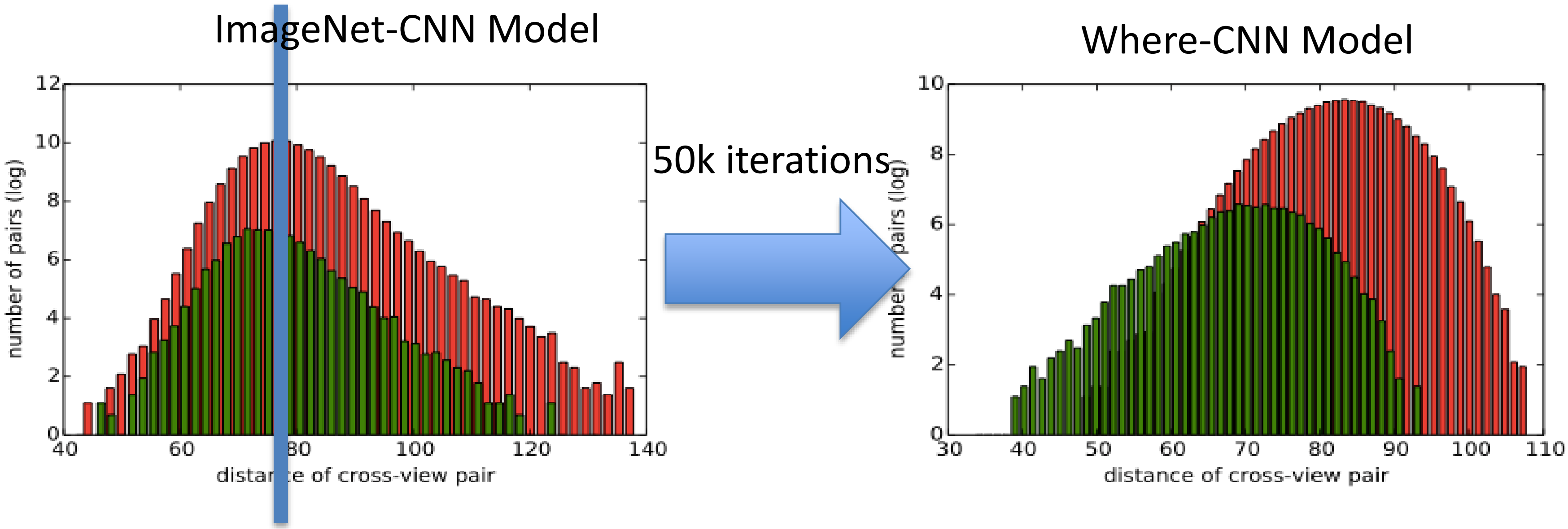


Hadsell, Chopra, Yann LeCun, **Dimensionality Reduction by Learning an Invariant Mapping,** CVPR06

Pair Distance Distribution

Margin

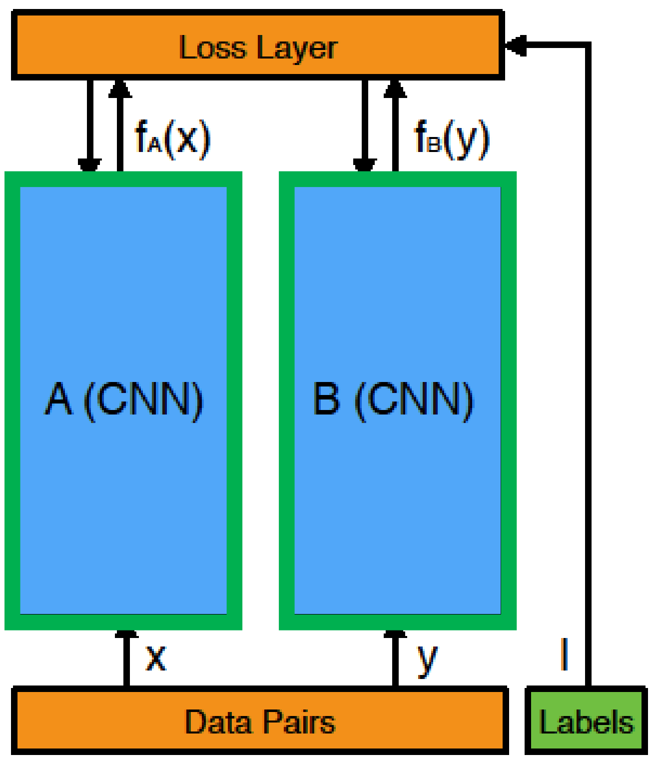
Green: positive pairs
Red: negative pairs



Quantitative Evaluation (AP)

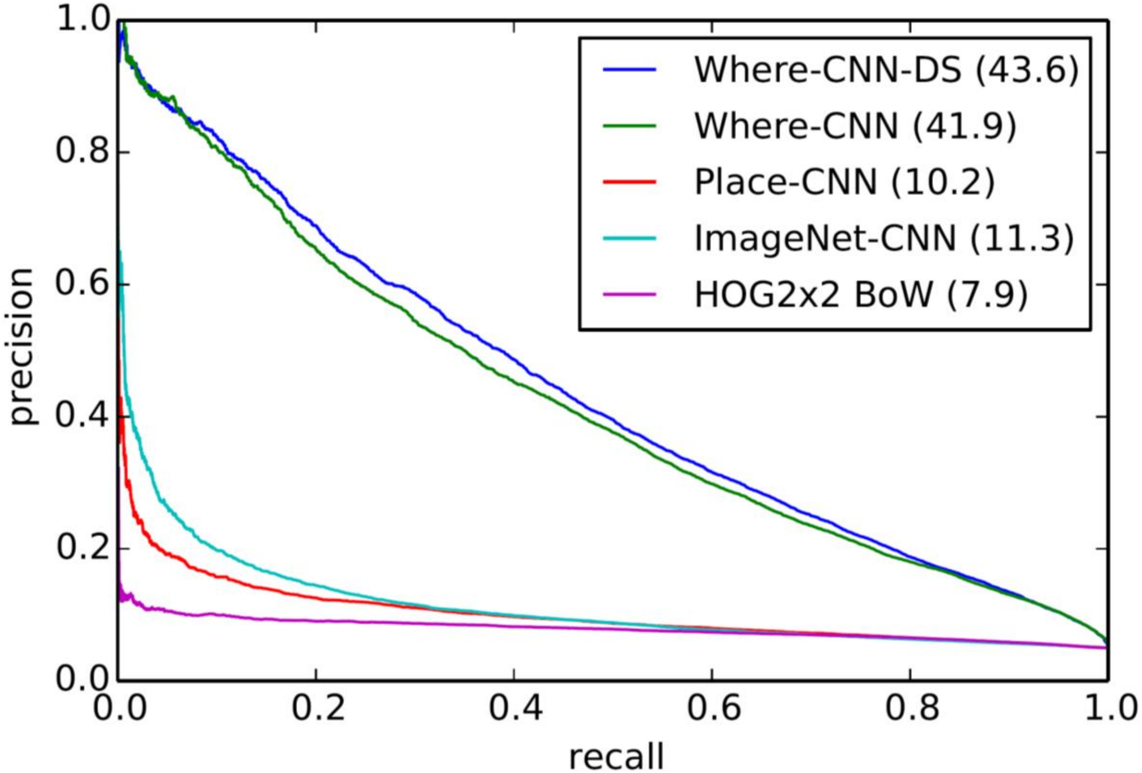
- Random: 5% (1:20 pos. to neg. pairs)
- HoG2x2 (BoW): 7.9%
- Places-CNN: 10.2%
- ImageNet-CNN: 11.3%
- **Where-CNN (ours): 41.9%**

Share The Same Parameters?

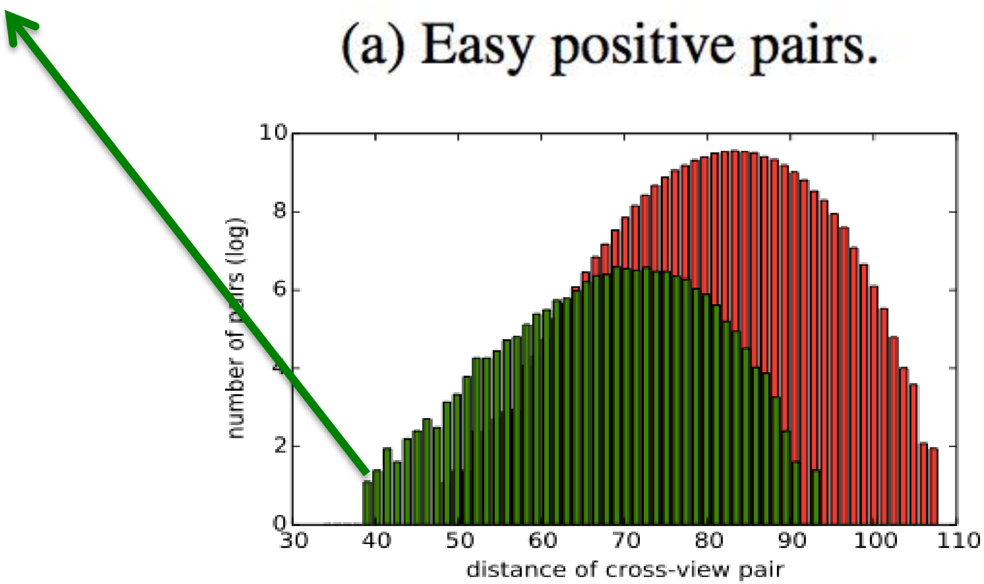


- For face verification A and B share parameters
- For ground-aerial image pairs, should A, B share parameters?

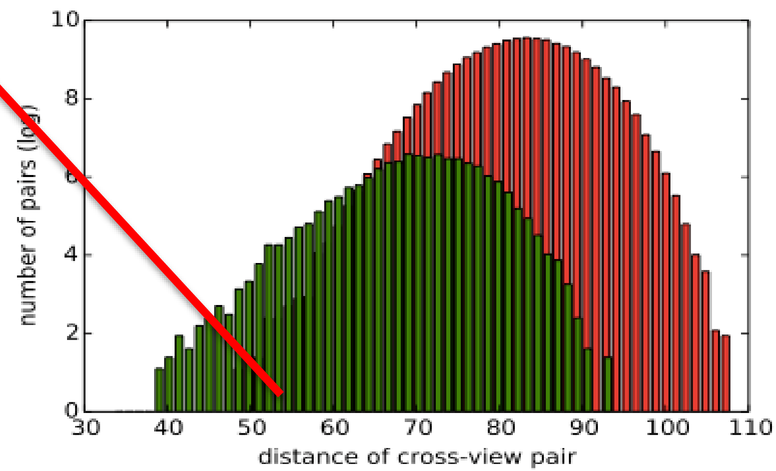
Quantitative Evaluation

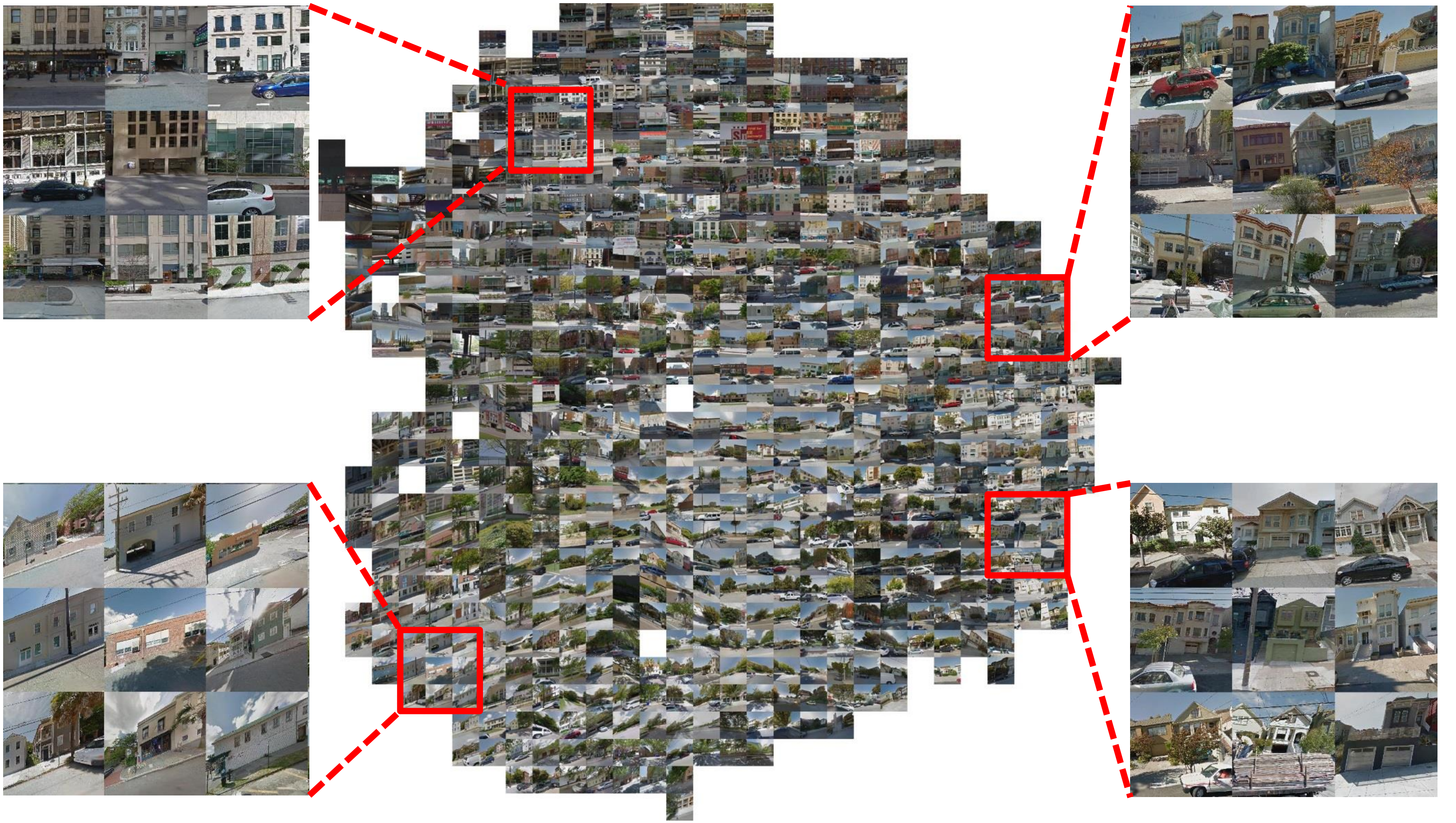


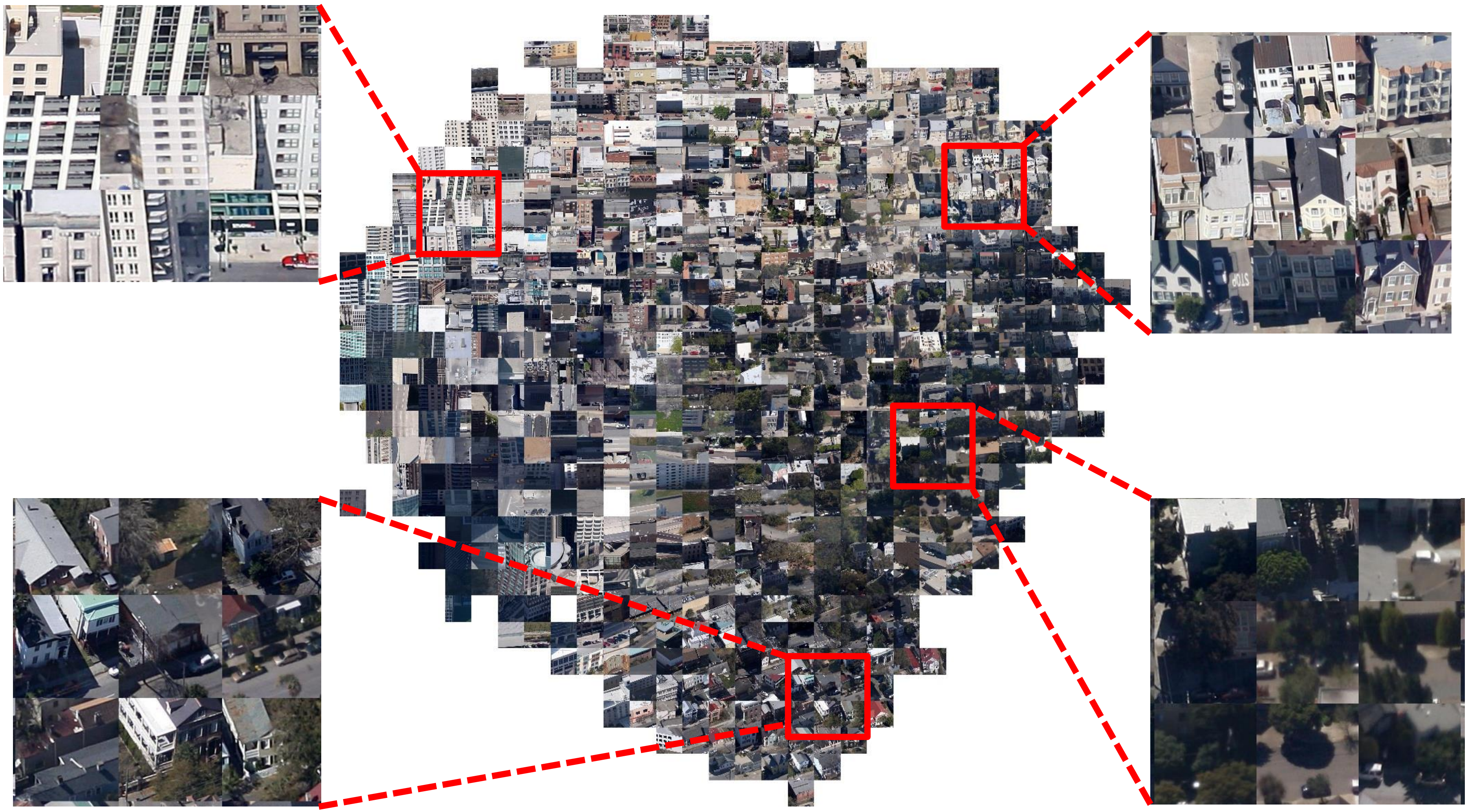
(a) Easy positive pairs.

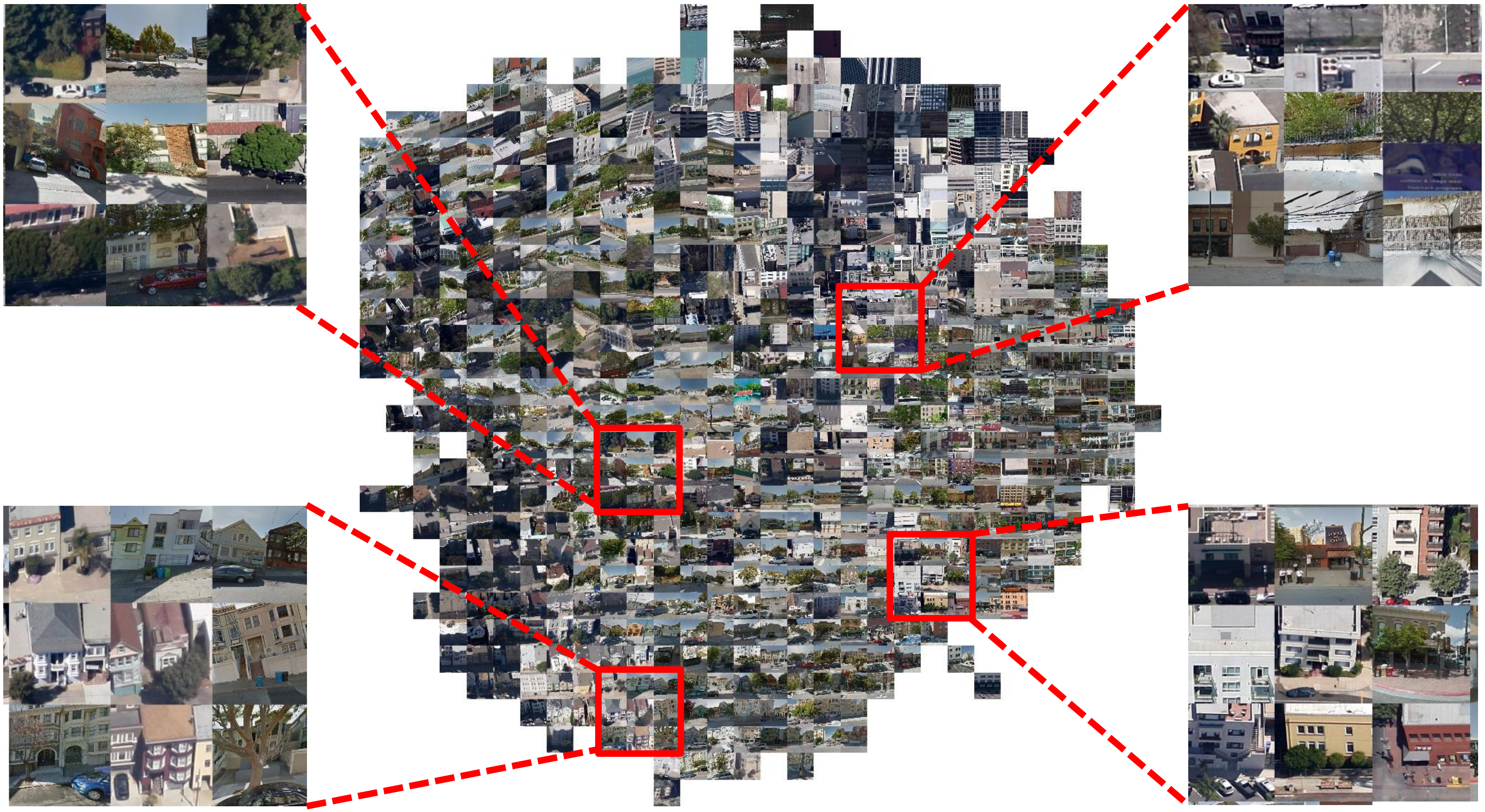


(b) Hard negative pairs.









Strongest Activations of Particular Units

Geolocalization

San Francisco

San Diego

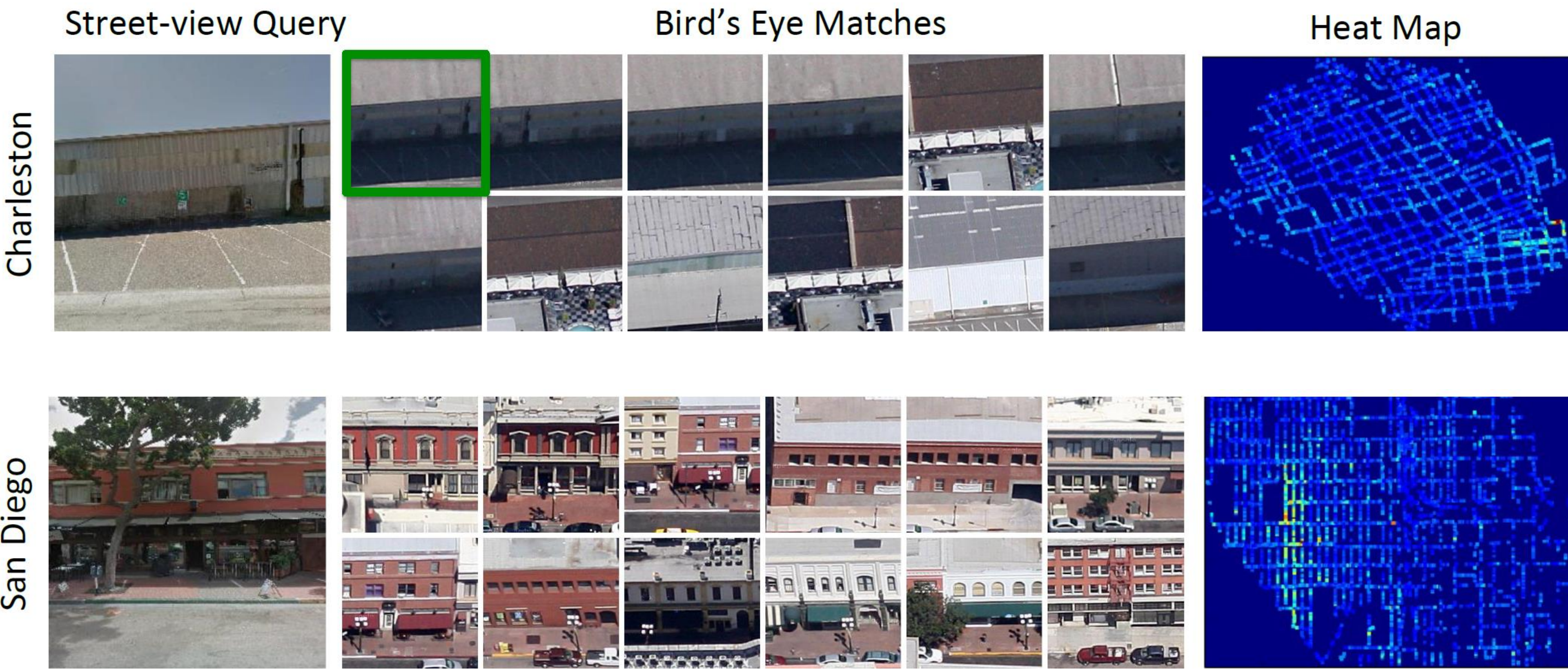
Chicago

Charleston

Tokyo

Rome

Lyon


Street-view Query
Bird's Eye Matches
Heat Map
Charleston
San Diego

Conclusions

- Localize images without corresponding ground-level images
- Create a large-scale training dataset from public data sources
- Learning feature representations for matching cross-view images