Learning Deep Representations for Ground-to-Aerial Geolocalization

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

† Cornell Tech § Brown University

Motivation

Image based Geolocalization

- ➢ Most previous methods: match query image to ground-level images with know locations.
- ➢ Most of the Earth does not have ground-level reference photos available. Fortunately, more complete coverage is provided by aerial imagery.
- ✓ Localize a ground-level query image by matching it to aerial imagery.

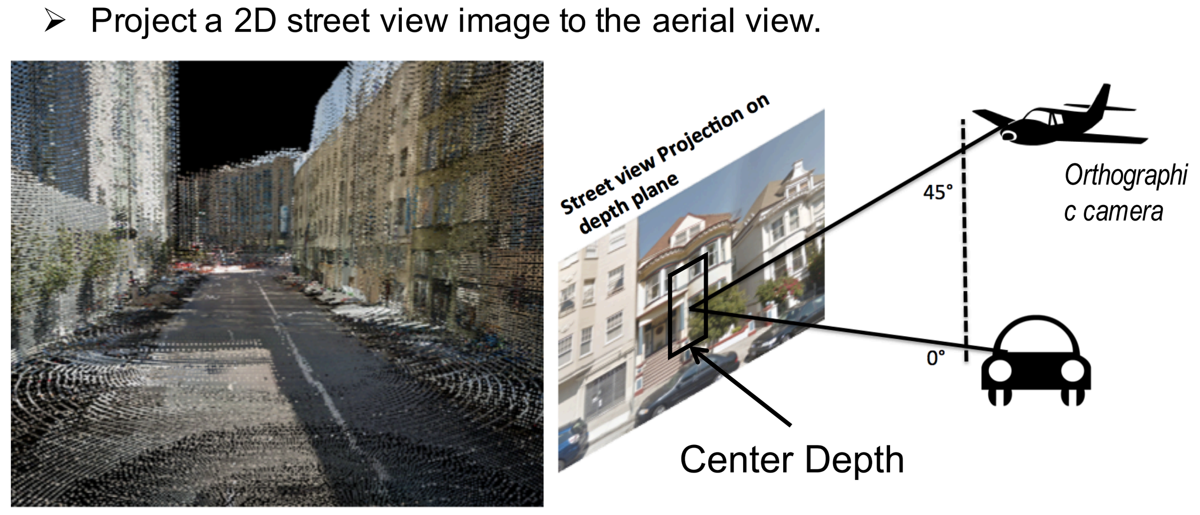
SIFT + RANSAC fails

- ➢ It is challenging to do key points matching from street-view to aerial-view
- ➢ Occlusions and differences in scale, capturing time, image resolution, etc.

Dataset Collection

Ground-to-Aerial Alignment for establishing ground-truth

- ➢ Known: street-view car heading direction; GPS location; depth estimates
- ➢ Project a 2D street view image to the aerial view.



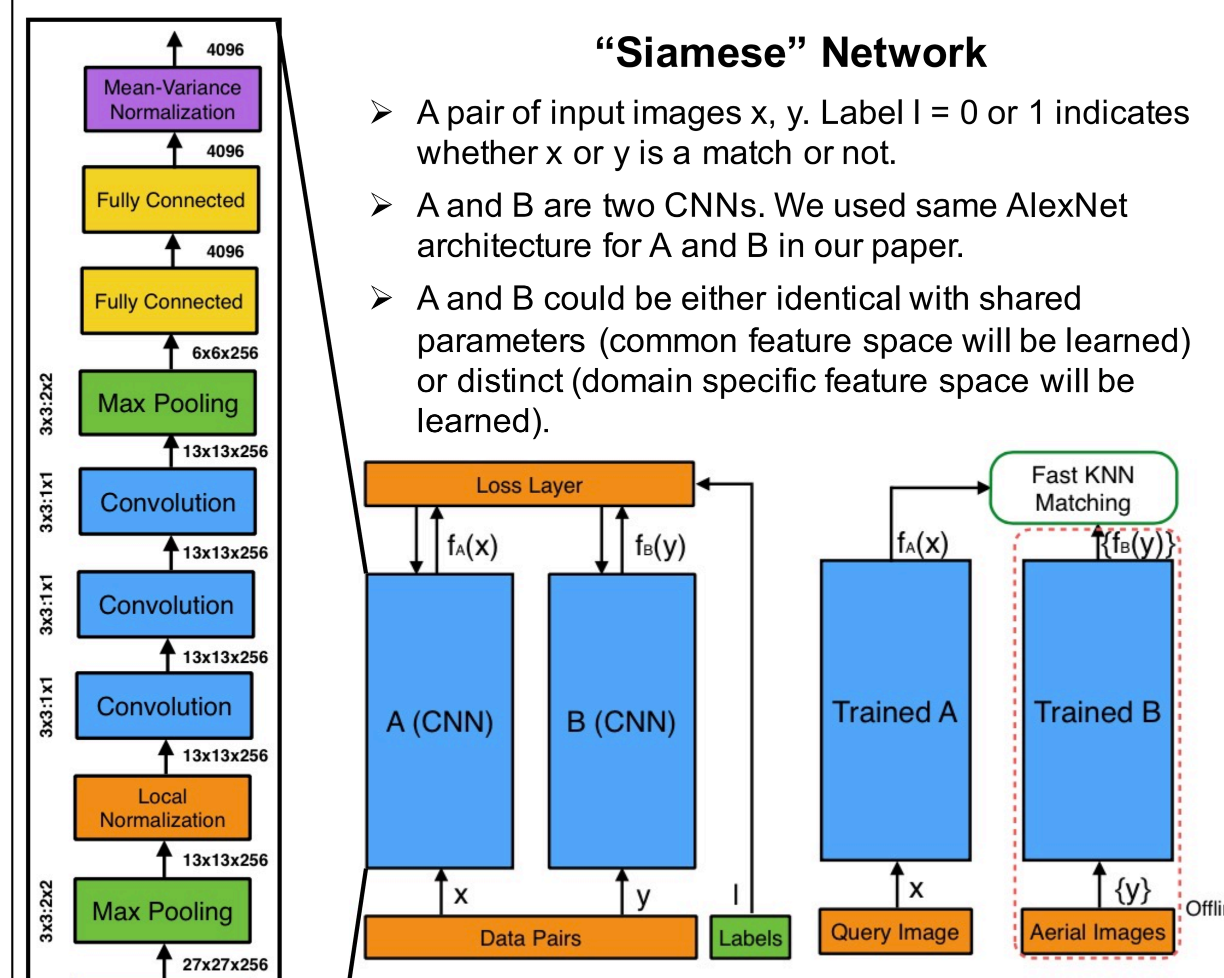
Dataset Statistics

- ➢ 7 cities (4 US and 3 non-US): San Francisco, San Diego, Chicago, Charleston, Tokyo, Rome, Lyon.
- ➢ 78K aligned street-view and aerial-view pairs.
- ➢ Image resolution: 15 x 15 meters (256 x 256 pixels)
- ➢ Cardinal viewing direction (azimuth) of 0° 90° 270° for training, 180° for testing.

Learning Deep Feature Embedding

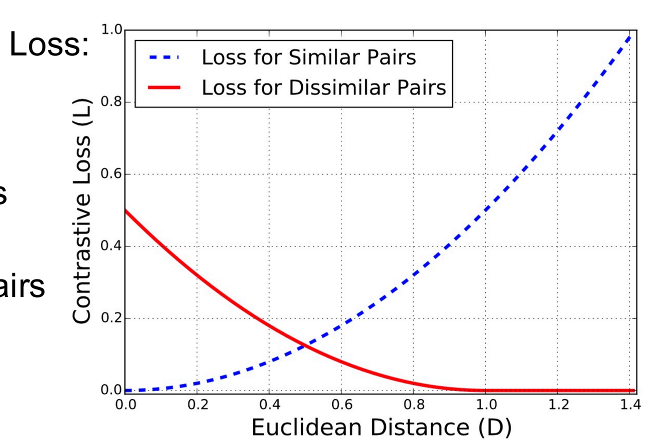
"Siamese" Network

- ➢ A pair of input images x, y. Label l = 0 or 1 indicates whether x or y is a match or not.
- ➢ A and B are two CNNs. We used same AlexNet architecture for A and B in our paper.
- ➢ A and B could be either identical with shared parameters (common feature space will be learned) or distinct (domain specific feature space will be learned).



(a) Training (b) Testing

- ➢ Contrastive Loss:
- • Pull together matched pairs
- • Push away unmatched pairs

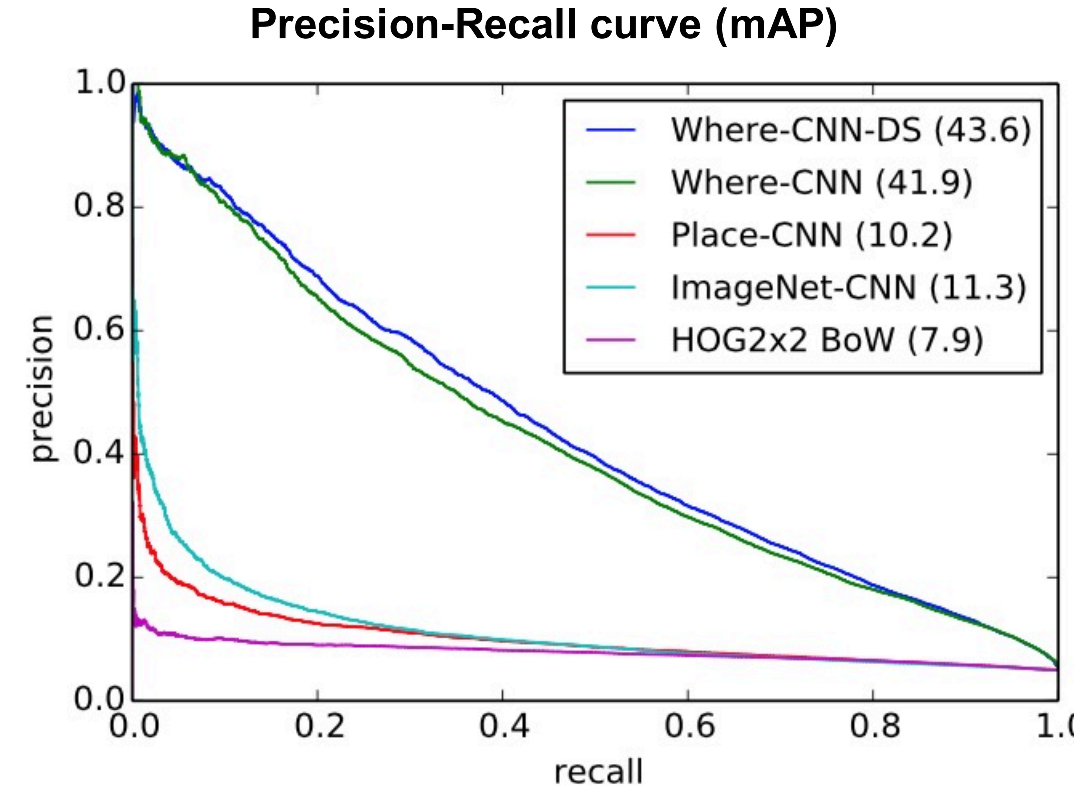


Experiments – Location Verification

Experimental Setting

- ➢ Location verification: given a pair of street-view and aerial-view images, identifying whether this pair comes from same location or not.
- ➢ 37.5K (12.5K) positive pairs, together with 20x more generated negative pairs from 4 US cities are used for training (testing). In total 0.79M (0.26M) pairs. Fine-tuned from pre-trained AlexNet on ImageNet.

Precision-Recall curve (mAP)

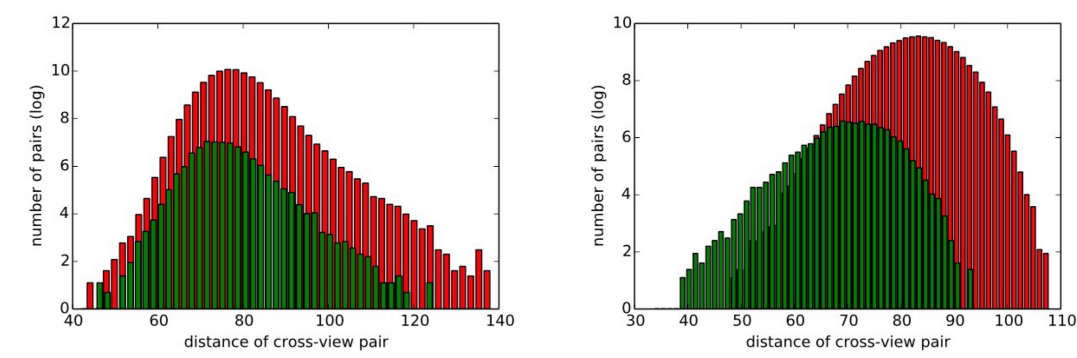


- ➢ Where-CNN is our CNN model; DS means domain specific; Places-CNN and ImageNet-CNN are AlexNet feature from 2nd last fully-connected layer (fc7) trained on Places and ImageNet datasets respectively.

Detailed Analysis

Effectiveness of training

- ➢ Histogram of Pairwise distances on test set:



(a) ImageNet-CNN feature. (b) Where-CNN feature.

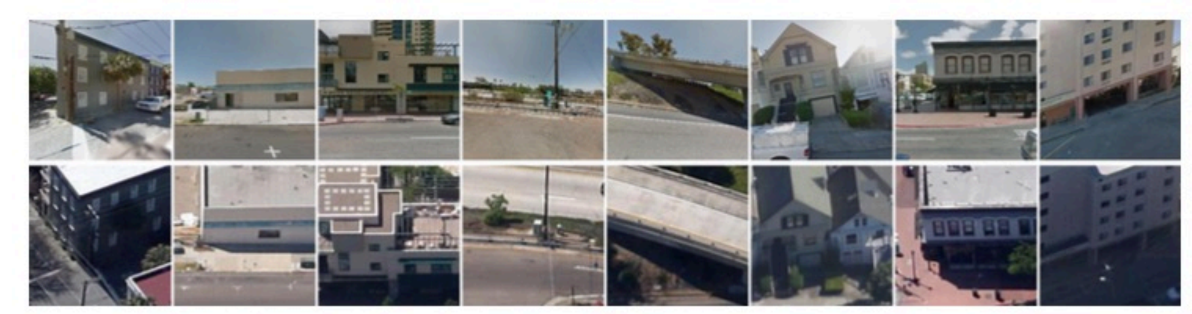
Robustness of initialization (fine-tuning)

- ➢ We fine-tuned our CNN from ImageNet and Places datasets:

Where-CNN	ImageNet Init.	Places Init.
AP	41.9%	41.4%

Easy positives and hard negatives

- ➢ The most similar true positives and false positives matches on test set.

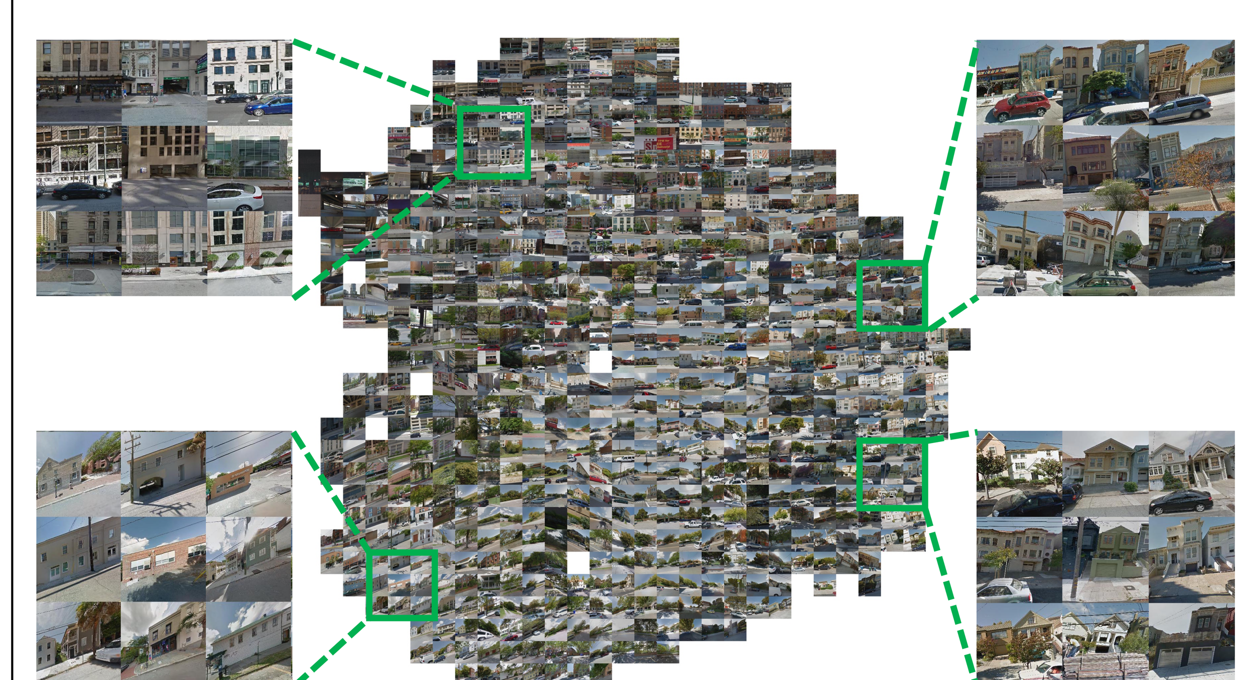


(a) Easy positive pairs.

(b) Hard negative pairs.

2-Dimentional Feature Embedding for Street-view

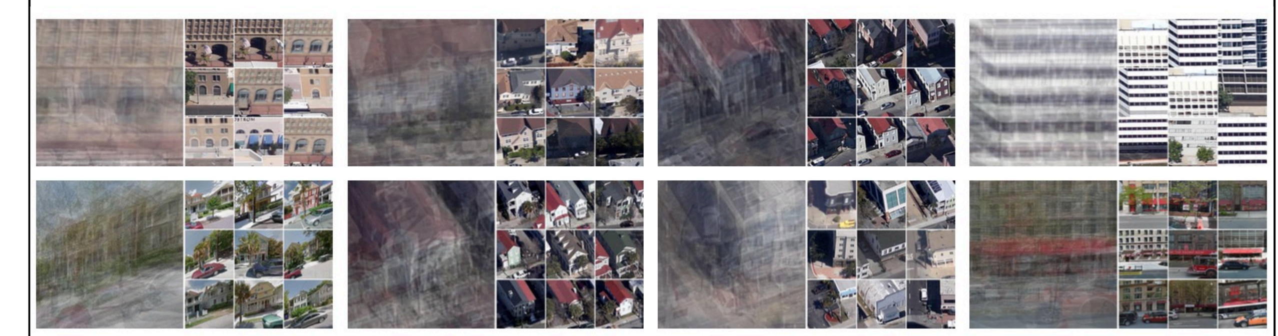
- ➢ We extracted 4096 dimensional features from Where-CNN on test set and used t-SNE for dimension reduction.



Visualization of Units' Receptive Fields

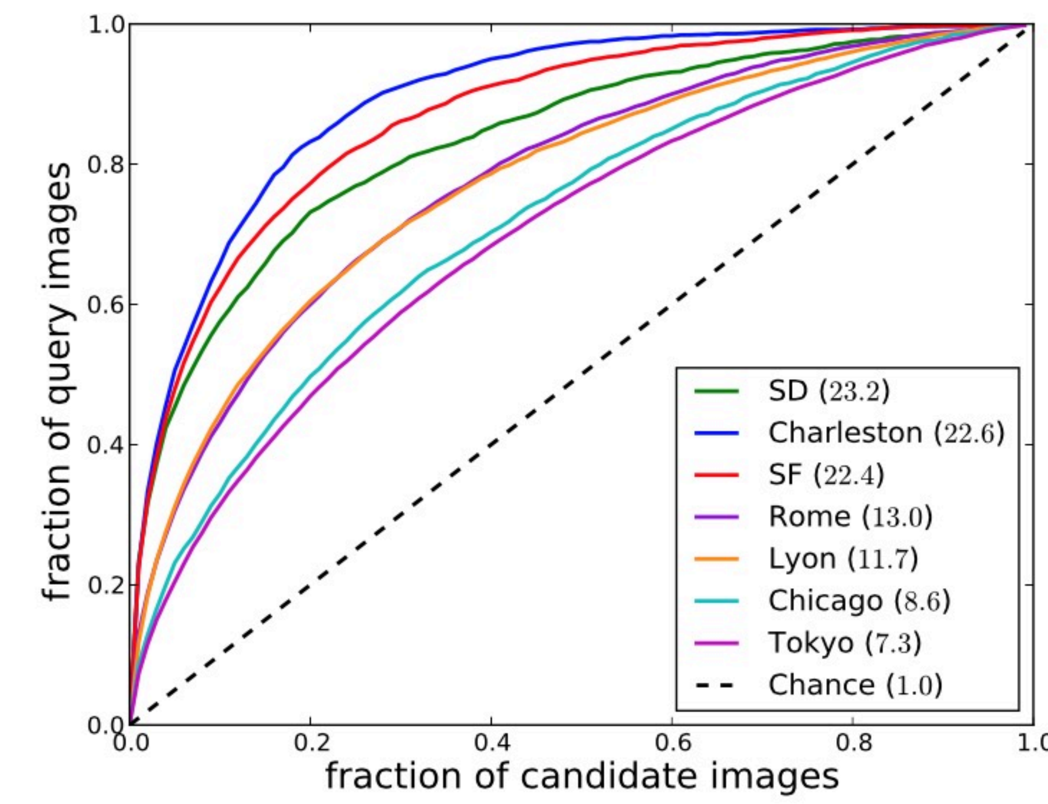
Strongest Activations of Particular Units

- ➢ Illustration of the average images and the top 9 images that activate a certain unit most strongly at the output feature layer (fc7 layer).



Cross-city Geolocalization

- ➢ Trained Where-CNN on 4 US cities and test on 3 novel non-US cities.
- ➢ Fraction of queries with true match in top 1% nearest neighbors.
- ➢ x-axis: fraction of nearest neighbors considered; y-axis: fraction of queries with true match in the nearest neighbors considered.



Geolocalization Examples

- ➢ Examples of query images, the top 12 matched aerial images for that query, and the heat map that indicates possible locations.
- ➢ The first 2 rows are success cases; and last 2 rows are failure cases.