

Introduction to General Visual Geo-Localization

Han-Pang Chiu

Center for Vision Technologies
SRI International, Princeton, NJ, USA
Email: han-pang.chiu@sri.com

June 18, 2023



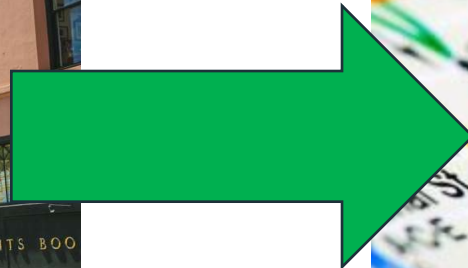
SRI International[®]

Outline

- **Image-Based Geo-Localization**
- Coarse Search: Approaches and Comparison
- Coarse Search: Trends
- Fine Alignment
- Use Cases
- Conclusion

Image-based Geo-Localization

Goal: Estimate the 3D geodetic position (or 3D pose – including both position and orientation) based on a Query Image



Impact

Historical Imagery



Improve GPS Accuracy



Personal Photo Album



GPS Denied/Challenged Environments



Image-Based Visual Geo-Localization – Coarse Search

Predicting the location of a query image with a database of geo-referenced data, which is also called Place Recognition in some fields.

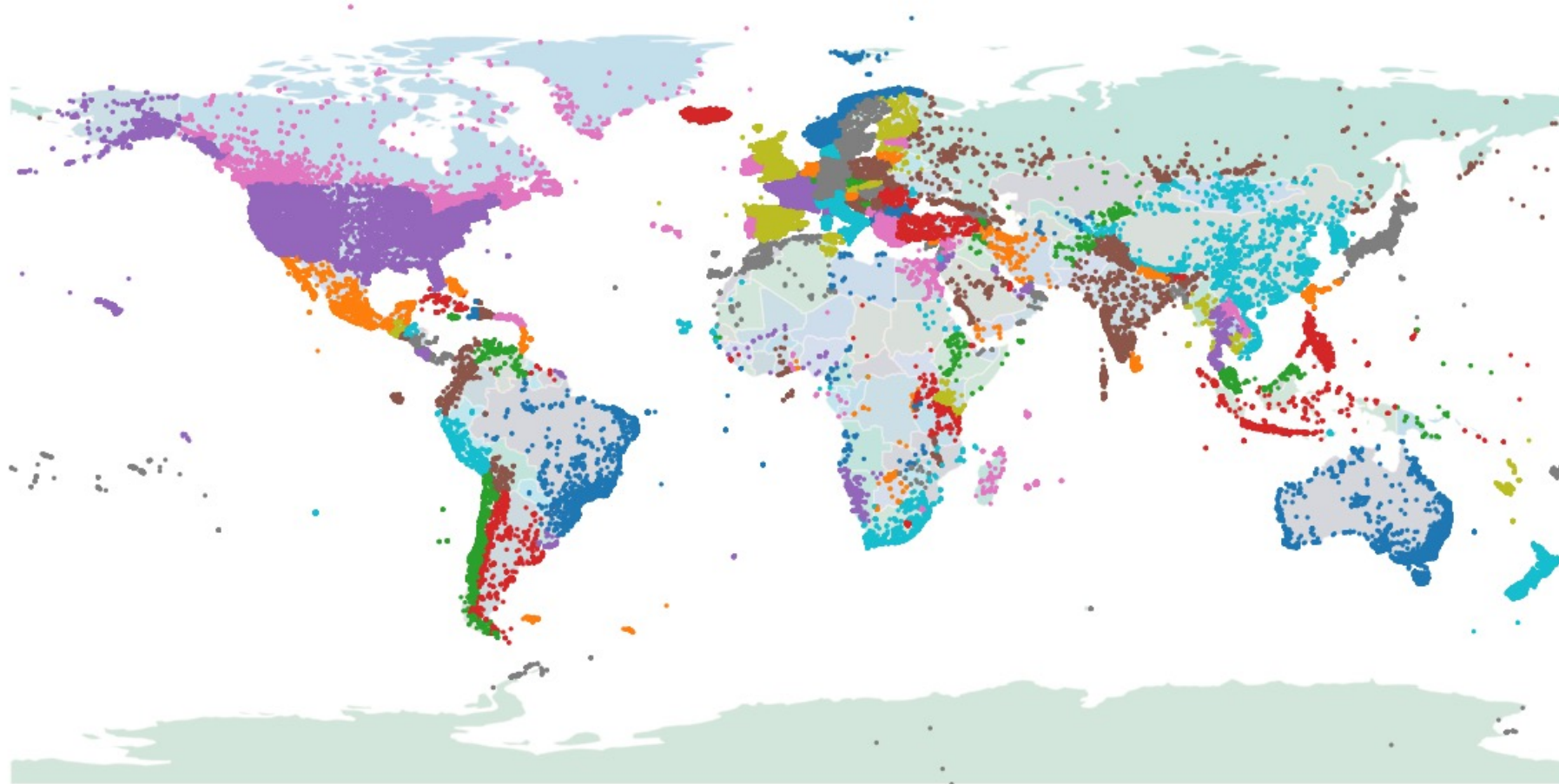
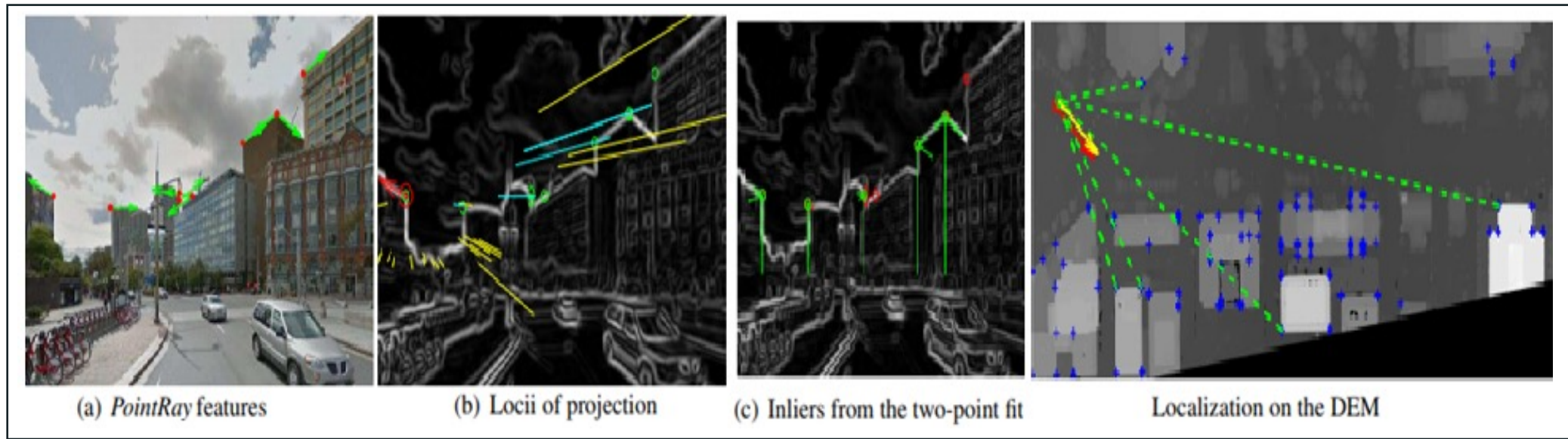


Image-Based Visual Geo-Localization – Fine Alignment

Given an initial 3D pose (from coarse search), registering the query image to the 3D geo-referenced data to further refine the 3D pose of this query image.

- It requires detailed 3D information in the database (such as 3D point cloud).
- It is also called geo-registration.

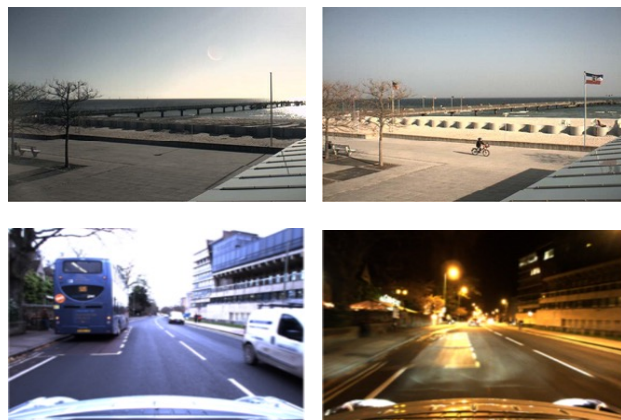


M. Bansal et al., “Geometric Urban Geo-Localization”, CVPR 2014.

Image-Based Visual Geo-Localization

Cross-Time, Cross-View, and Cross-Modal

Cross-Time



Sample Pairs (Ground RGB)

Cross-View



Sample Pairs (Ground-Aerial RGB)

Cross-Modal



Sample Pairs (Ground RGB-OpenStreetMap)

Low

Availability of Geo-Referenced Database

High

Difficulty in Image-Based Visual Localization

Challenges

The real world is dynamic with lots of moving objects and continuous scene changes.

Visual geo-localization methods need to be adaptive to changes in season, time of day, clutter in the scene etc.

How to scale the method for real world-wide applications?

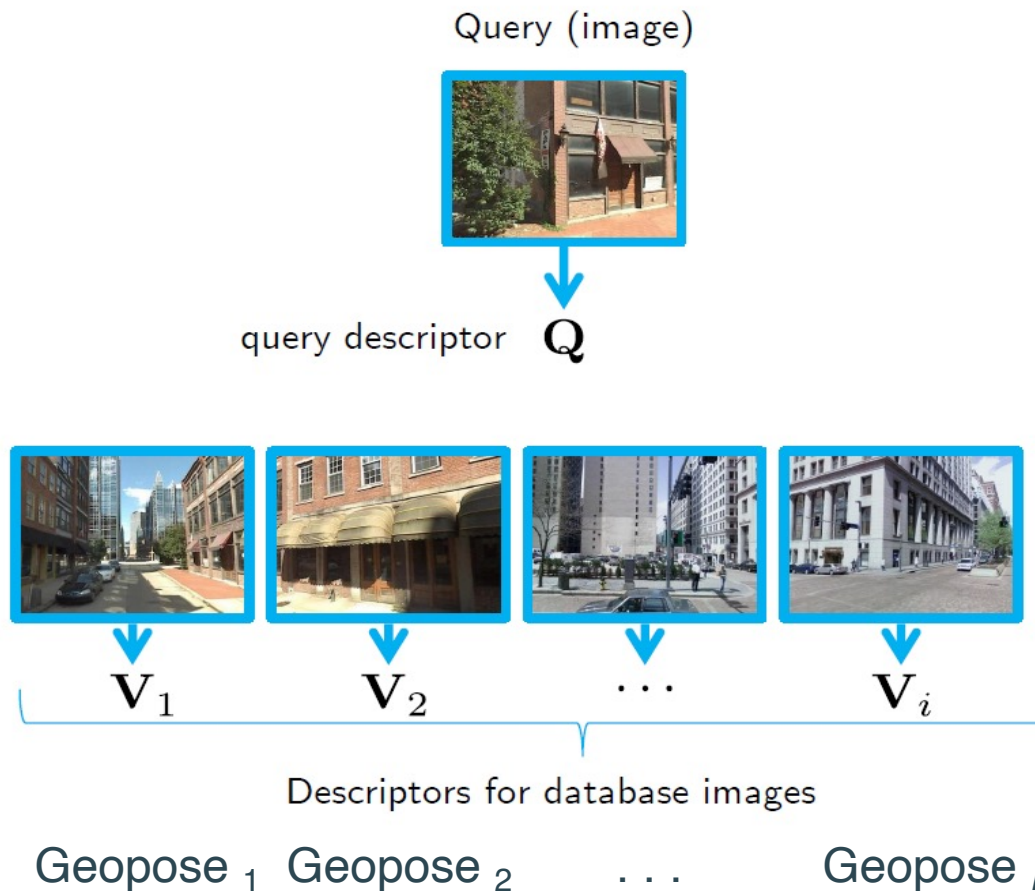


Outline

- Image-Based Geo-Localization
- **Coarse Search: Approaches and Comparison**
- Coarse Search: Trends
- Fine Alignment
- Use Cases
- Conclusion

Coarse Search: Retrieval-Based Geo-Localization

Matching a query image to a geo-referenced databases, which is also called database retrieval in some fields.

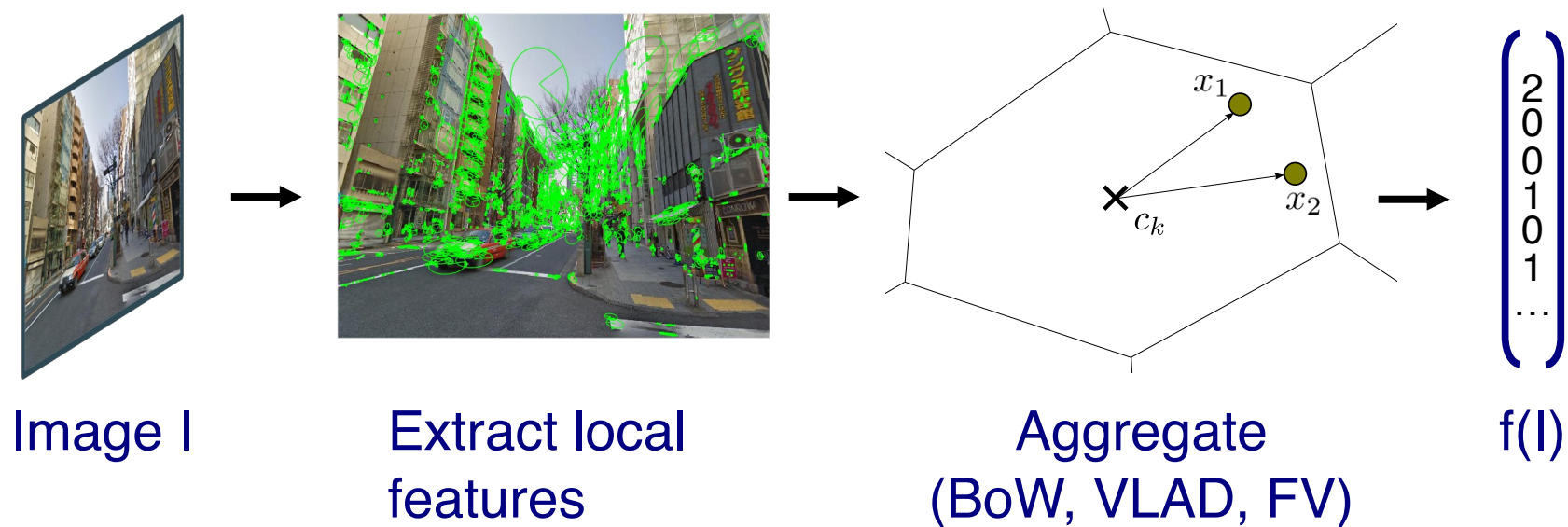


$$i^* = \arg \max_{i \in 1 \dots N} s(Q, V_i)$$

Nearest Neighbor search

Image Representation in Retrieval-Based Geo-Localization

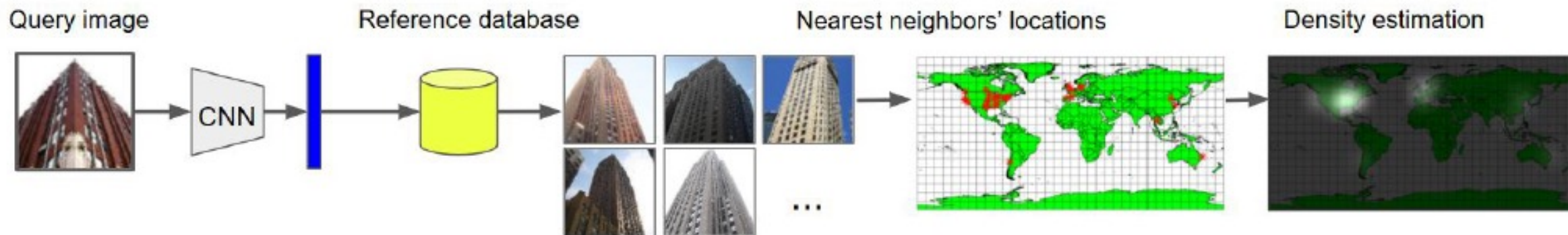
- Local Descriptors: SIFT, SURF, BRIEF etc.
- Aggregation of local features: Bag-of-Words (BoW), VLAD, FV etc
- Global Descriptors: Encode holistic properties of the scene, such as GIST
- Deep Learned Representation: NetVLAD [1] etc



[1] NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

Representative Work: IM2GPS [1] & IM2GPS-Deep [2]

- IM2GPS was the first attempt of applying image retrieval for world-scale geo-localization problem.
- IM2GPS-Deep greatly improves IM2GPS results by using deep learning models to extract features from the image.



[1] Hays, James, and Alexei A. Efros. "IM2GPS: estimating geographic information from a single image." CVPR, 2008.

[2] Vo, Nam, Nathan Jacobs, and James Hays. "Revisiting im2gps in the deep learning era." ICCV. 2017.

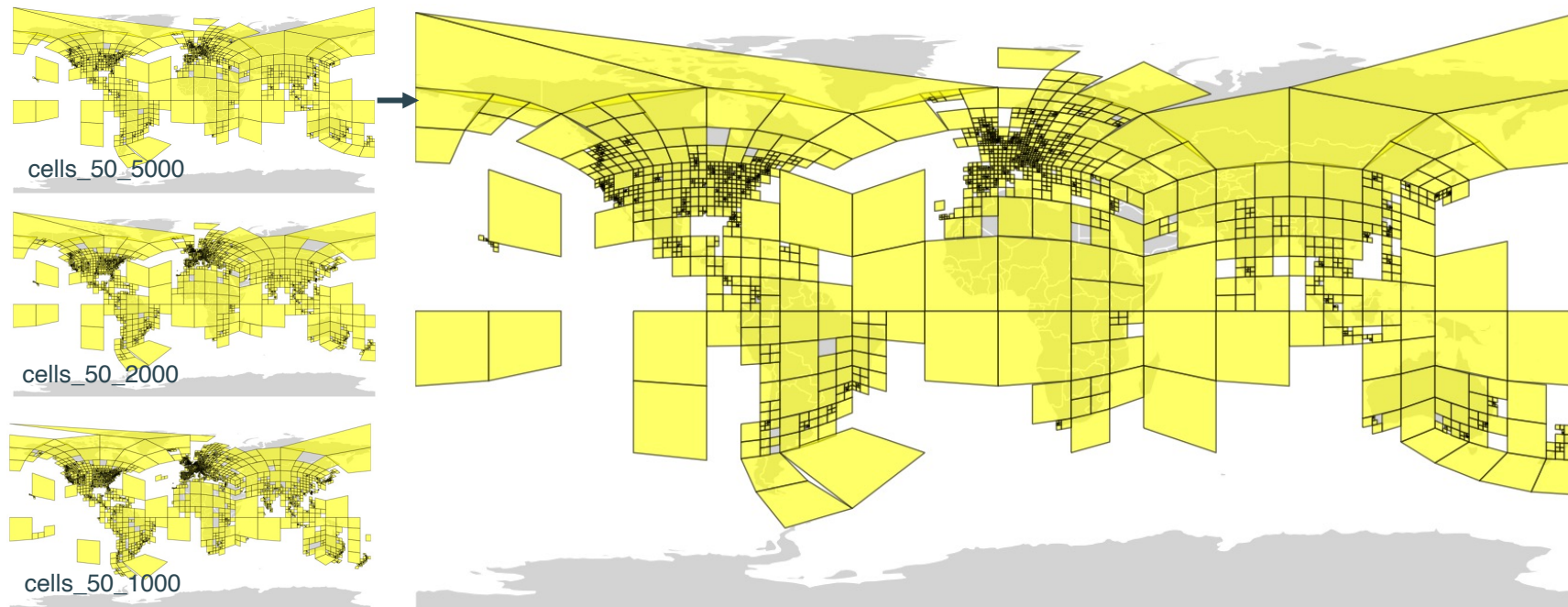
***For recent development in retrieval-based approach, please attend our 9am talk - Recent Image Geo-localization Benchmark and Large-scale Real-world Scenarios (Carlo Masone /Gabriele Berton)**

C. Masone et al, "A survey on deep visual place recognition", IEEE Access 2021

Coarse Search: Classification-Based Geo-Localization

Formulate geo-localization as a classification task on worldwide grids.

- (1) The world map is partitioned into grids (such as using Google's S2 library*).
- (2) A classification model assigns a query image into one of the grids.
- (3) The predicted geo-location of a query image is the center of the predicted grid.



*<https://code.google.com/archive/p/s2-geometry-library/>

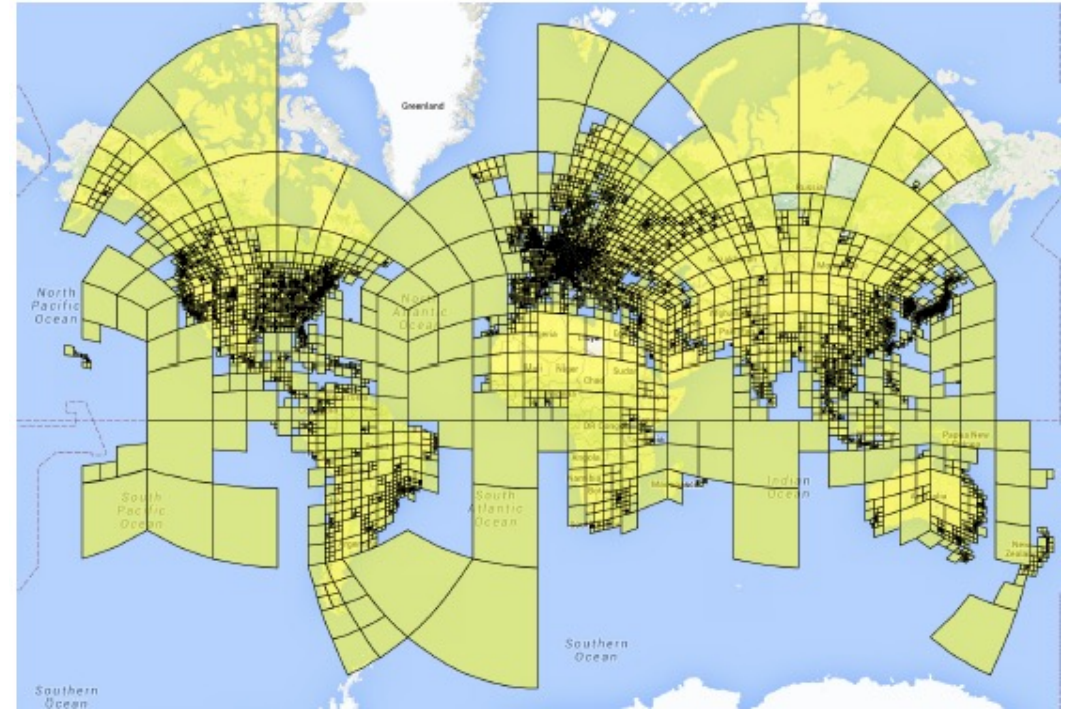
Representative Work: PlaNet [1]

- PlaNet is the first implementation for classification-based geo-localization.

- (1) The world is initially partitioned into 6 cells.
- (2) Recursively subdivide cells until no cell contains more than a certain fixed number t_1 of photos.
- (3) Discard all cells containing less than a minimum of t_2 photos ($t_1 = 10,000$ and $t_2 = 50$)

Advantage: (i) training classes are more balanced, (ii) effective use of the parameter space because more model capacity is spent on densely populated areas, (iii) the model can reach up to street-level accuracy in city areas where cells are small.

Classification Model: Train a CNN based on the Inception architecture [2] with batch normalization [3]. The SoftMax output layer has one output for each cell in the partitioning.



Adaptive partitioning of the world into 26,263 S2 cells.

[1] Weyand, Tobias, Ilya Kostrikov, and James Philbin. "Planet-photo geolocation with convolutional neural networks." ECCV 2016.

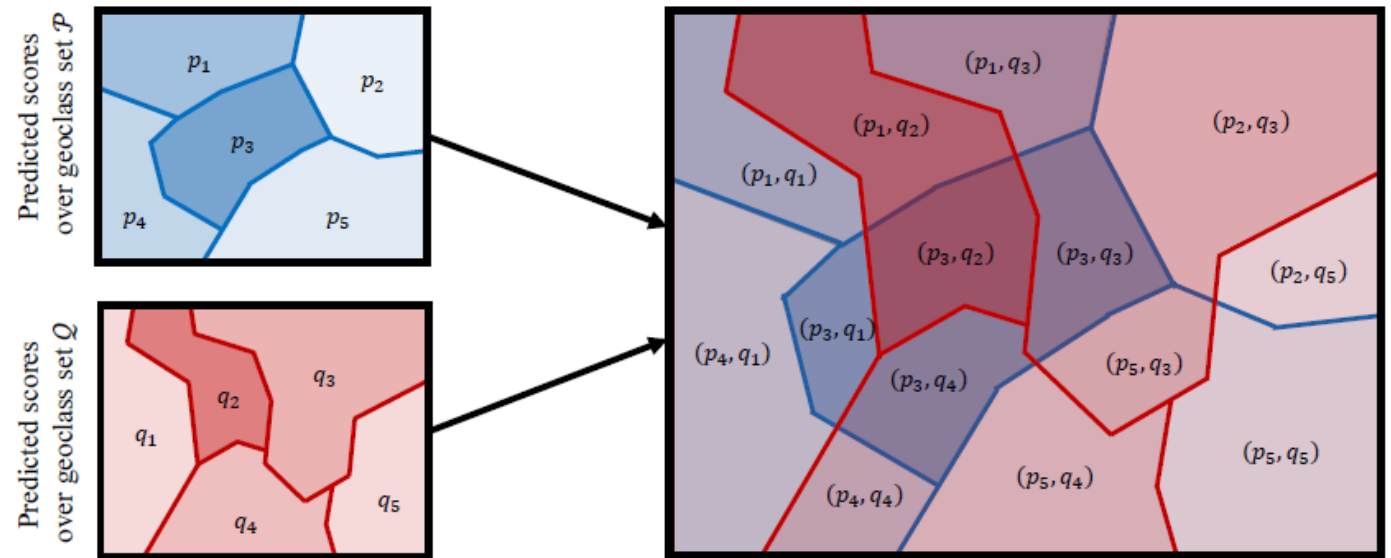
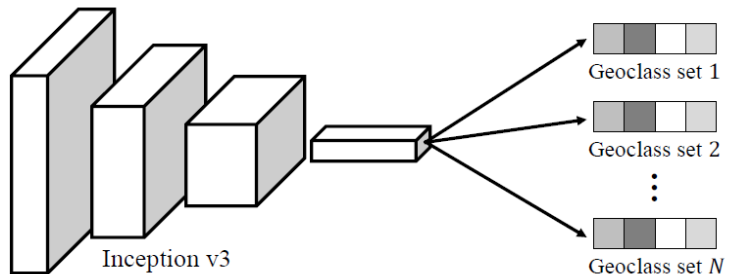
[2] C. Szegedy et al.. Going Deeper with Convolutions. In CVPR, 2015.

[3] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In ICML, 2015.

Representative Work: CPlaNet [1]

- Use a combinatorial partitioning strategy to find the balance between accuracy and efficiency.
- Increasing the number of classes (reducing the area of the region corresponding to each class) increases the number of parameters while decreasing the size of the training dataset per class.

Advantage: (i) Fine-grained classes with fewer parameters: $2 \cdot (5 \cdot F) < 14 \cdot F$, (2) More training data per class, (3) More reasonable class sets
Classification Model:



Visualization of combinatorial partitioning:
Two sets with 5 geoclasses form 14 distinct classes

Inference: (1) Summarize normalized score from each classifier (set) on each distinct cell.
(2) Predict the location as average position of training images from the highest score cell.

[1] Seo, Paul Hongsuck, Tobias Weyand, Jack Sim, and Bohyung Han. "Cplanet: Enhancing image geolocation by combinatorial partitioning of maps." ECCV. 2018.

Comparison: Retrieval vs Classification

	Retrieval	Classification
Fine Scale	Better at finer scale modeling (resolution can be controlled by training images)	Finer cells increase the number of parameters in models, and reduce the number of training images in each cell
World-Wide Modeling	Expensive in space and time for large-scale applications (data collection and database storage of feature vectors for all reference images)	Easier to move to world-wide large-scale applications due to its formulation
Environment	Require the database has significant overlap with the query image. May limit to specific landmarks in images	May generalize better to different environments (urban cities and rural scenes)

*Combining classification and retrieval can be a feasible direction to improve both generalization and accuracy.

***For newest work in coarse search, please attend our 10:45am talk - R2Former: Unified Retrieval and Reranking Transformer for Place Recognition (Sijie Zhu) and our 11:15am talk - Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes (Mubarak Shah).**

Outline

- Image-Based Geo-Localization
- Coarse Search: Approaches and Comparison
- **Coarse Search: Trends**
- Fine Alignment
- Use Cases
- Conclusion

Recent Visual Geolocation Results

Table 1: City-level (25 km) Visual Geolocation Accuracy (a_r)

Visual Geolocation Accuracy

$$a_r \equiv \frac{1}{N} \sum_{i=1}^N u \left(d(l_{gt}^{(i)}, l_{pred}^{(i)}) < r \right)$$

d - Great Circle Distance

l_{gt} - ground-truth location

l_{pred} - predicted location

u - indicator function

r - radius

Benchmark Datasets

Year	Architecture	IM2GPS	IM2GPS3k	YFCC4k	YFCC26k	Place Plus 2.0
2018	ISNs [1]	43.0	28.0	16.5	12.3	-
2018	CPlaNet [2]	37.1	26.5	14.8	11.0	-
2021	MvMF [3]	44.7	29.8	14.4	-	-
2022	Translocator [4]	48.1	31.1	18.6	17.8	-
2022	IM2City [5]	-	32.1	-	-	85.92
2022	SemP [6]	42.6	31.4	22.3	-	-

1. Muller-Budack, Eric, Kader Pustu-Iren, and Ralph Ewerth. "Geolocation estimation of photos using a hierarchical model and scene classification." *ECCV*. 2018.
2. Seo, Paul Hongsuck, et al. "Cplanet: Enhancing image geolocation by combinatorial partitioning of maps." *ECCV*. 2018.
3. Izbicki, Mike, Evangelos E. Papalexakis, and Vassilis J. Tsotras. "Exploiting the earth's spherical geometry to geolocate images." *ECML PKDD*. Springer, Cham, 2020.
4. Pramanick, Shraman, et al. "Where in the World is this Image? Transformer-based Geo-localization in the Wild." *ECCV*. Springer, Cham, 2022.
5. Wu, Meiliu, and Qunying Huang. "IM2City: image geo-localization via multi-modal learning." *ACM SIGSPATIAL*. 2022.
6. Theiner, Jonas, Eric Müller-Budack, and Ralph Ewerth. "Interpretable Semantic Photo Geolocation." *WACV*. 2022.

#1: Incorporation of Scene Types

- There is huge diversity caused by various environmental settings, which requires specific features to distinguish different locations.
- Image-based geo-localization can be benefit from contextual knowledge about the environmental scene, since the diversity in the data space could be drastically reduced.
- Train and use different networks to handle different scene types can better address this diversity problem.

[1] Muller-Budack, Eric, Kader Pustu-Iren, and Ralph Ewerth. "Geolocation estimation of photos using a hierarchical model and scene classification." *ECCV*. 2018.

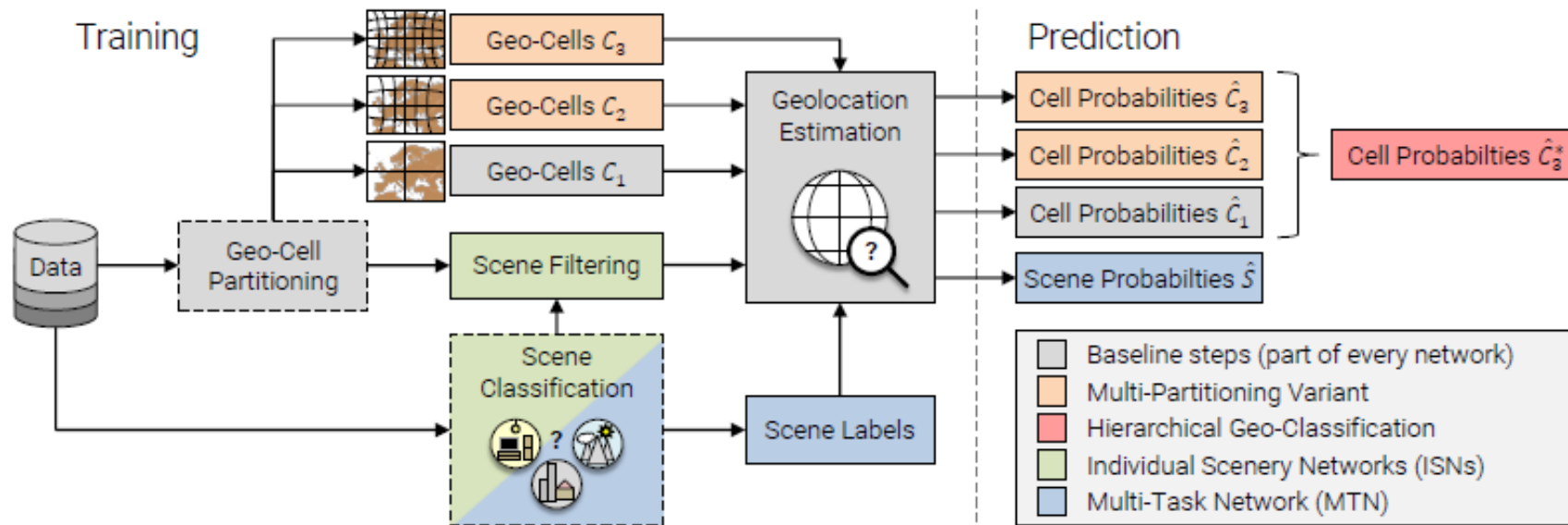
[2] Brandon Clark et al.. "Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes." *CVPR*. 2023.



Three main scene types [1]

Representative Work: ISN (Individual Scenery Network) [1]

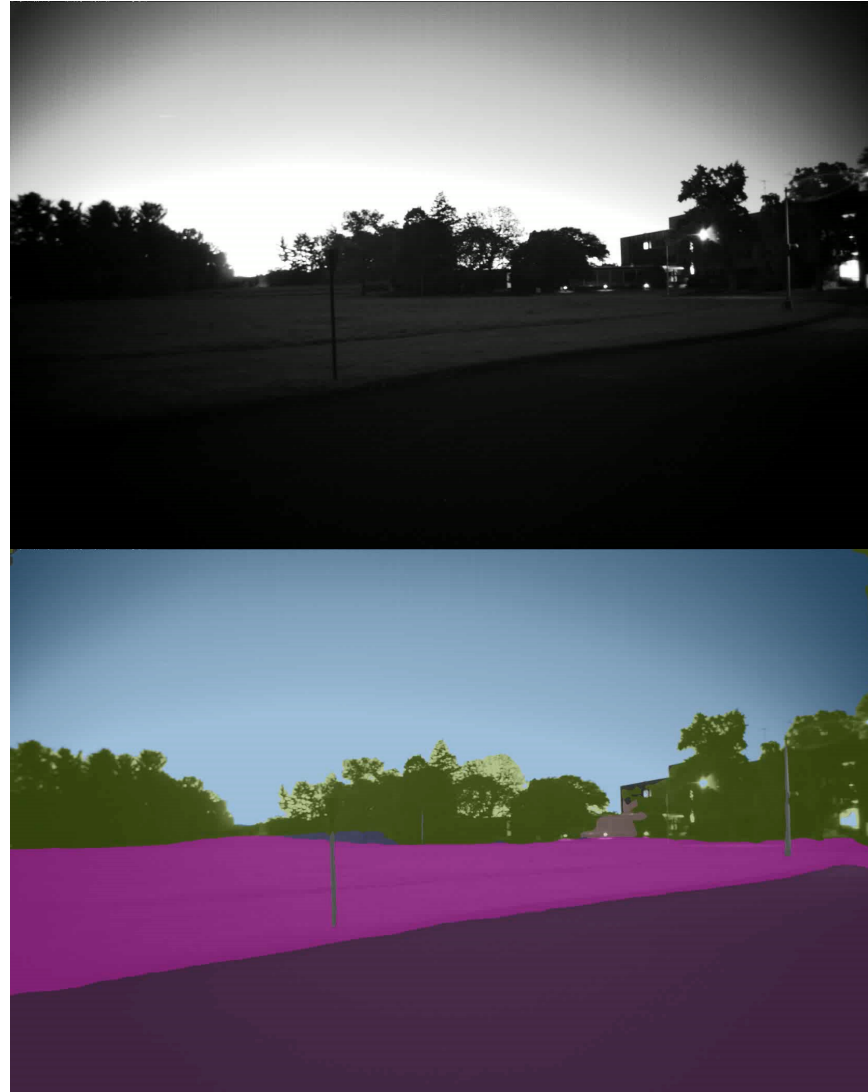
- Incorporating hierarchical knowledge at different spatial resolutions in a multi-partitioning approach.
 - Simultaneously learn geolocation estimation at multiple spatial resolutions using multi-partition loss function
- To classify scenes and extract scene labels, the ResNet model with 152 layers of the Places2 dataset is applied. The model has been trained on more than 16 million training images from 365 different place categories.
 - Three superordinate scene types (urban, natural, indoors) were used to train ISNs.
- Multi-Task Network (MTN) simultaneously train two classifiers for these complementary tasks.
 - Total loss of the MTN is defined by the sum of the geographical and scene loss.



[1] Muller-Budack, Eric, Kader Pustu-Iren, and Ralph Ewerth. "Geolocation estimation of photos using a hierarchical model and scene classification." *ECCV*, 2018.

#2: The Utilization of Scene Semantics

- The idea is to use both appearance and semantic cues to improve geo-estimation.
- Semantic segmentation guides the extraction of most informative visual elements for geo-localization.
 - Scene layout
 - Separation of objects (cars, people) from background and scene layout information.
 - Car and people should not contribute to geo-localization

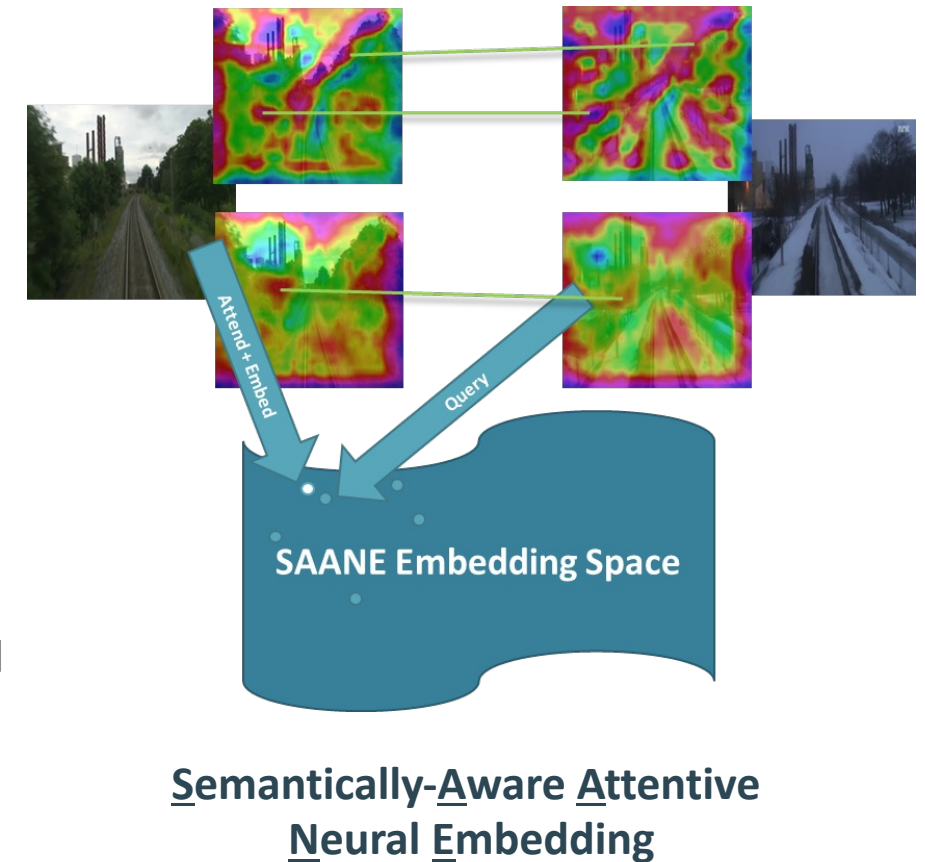


void	road
fence	pole
terrain	sky
sidewalk	building
traffic light	traffic sign
person	rider
wall	
vegetation	
car	

Semantic Color Palette

Representative Retrieval-Based Work: SAANE [1]

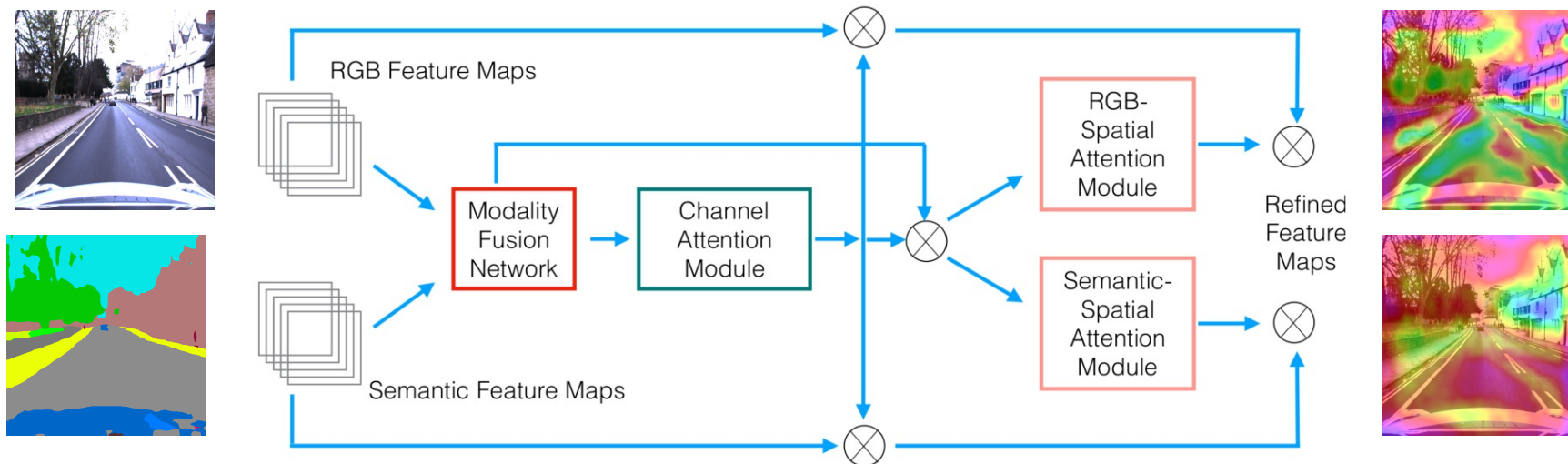
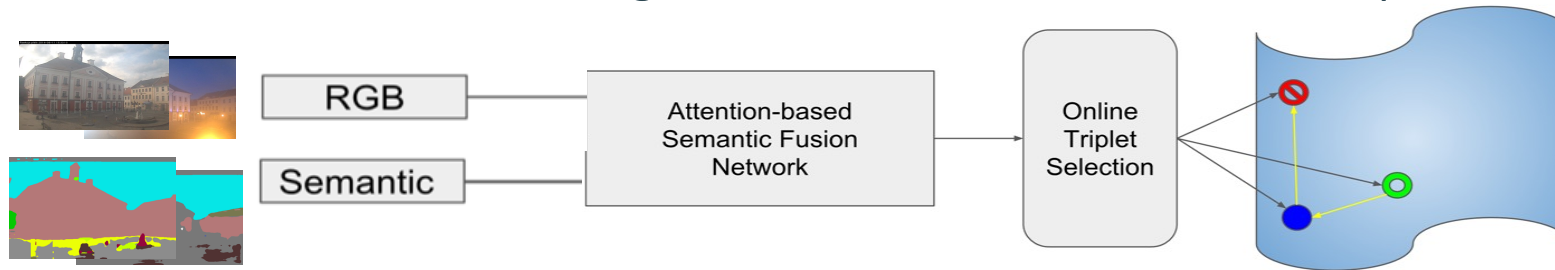
- **Goal:** Geo-tag a position for a given monocular **query image** by retrieval from a database of images of **known locations**, in the presence of large appearance changes across weather and illumination variations.
- SAANE uses embedding for this problem: A deep-learned compact Euclidean space where distances directly correspond to a measure of data similarity.
- **Semantic-Aware:** Our deep network model incorporates pixel-wise semantic features in learning the image embeddings.
- **Attention-Based:** We train self-attention modules to encourage our model to focus on semantically consistent spatial regions.
- Training data: ~2 million images collected from 2,685 static webcams.



[1] Zachary Seymour, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar, **Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization**. British Machine Vision Conference (BMVC), 2019

Representative Retrieval-Based Work: SAANE [1]

- The model incorporates pixel-wise semantics in learning the image embeddings.
 - Compared to low-level appearance descriptors, the spatial layout of semantic classes in the image yields scene descriptions that have a higher invariance to large changes in viewing conditions, to improve visual localization.
- The results show an 8–15% absolute improvement over our baseline from adding semantic information.
- Attention module combined with semantics gives an additional 4% absolute improvement on average.



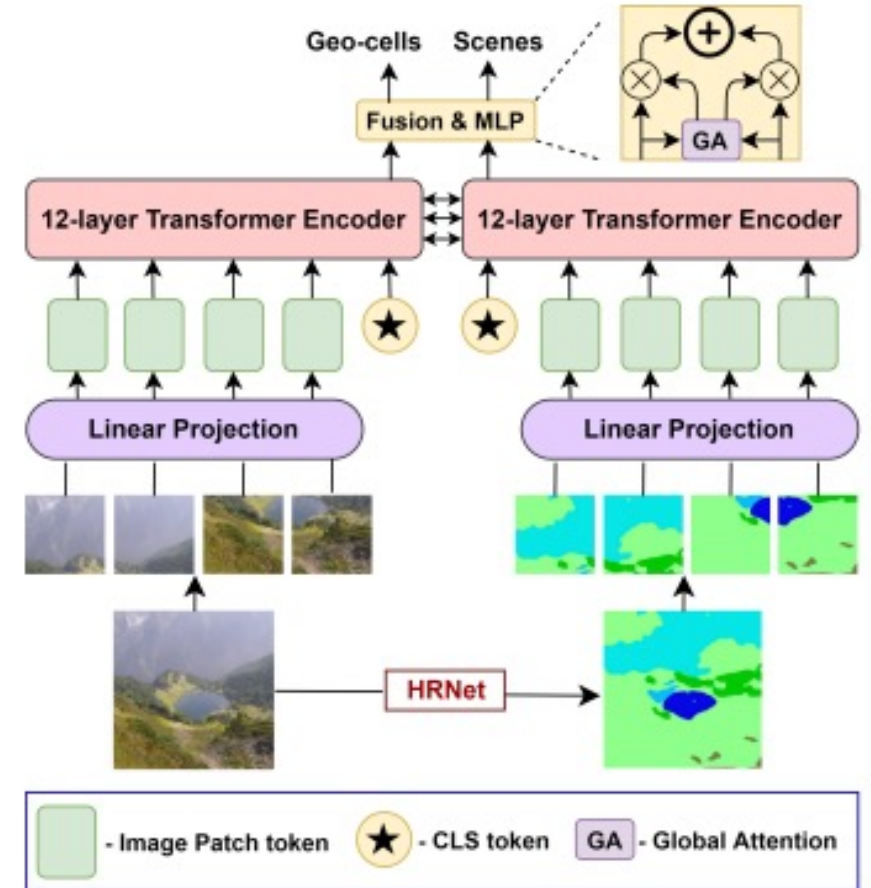
Demonstration: Localization Across Day & Night

- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.



Representative Classification-Based Work: Translocator [1]

- Use a vision transformer (ViT) [2] as the backbone.
 - An image is split into a sequence of 2-D patches which are then flattened and fed into the stacked transformer encoders through a trainable linear projection layer. An additional classification token (CLS) is added to the sequence, as in the original BERT approach [3].
 - A transformer encoder is made up with a sequence of blocks containing multi-headed self-attention (MSA) with a feed-forward network (FFN). FFN contains two multi-layer perceptron (MLP) layers with GELU non-linearity applied after the first layer. Layer normalization (LN) is applied before every block, and residual connections after every block.



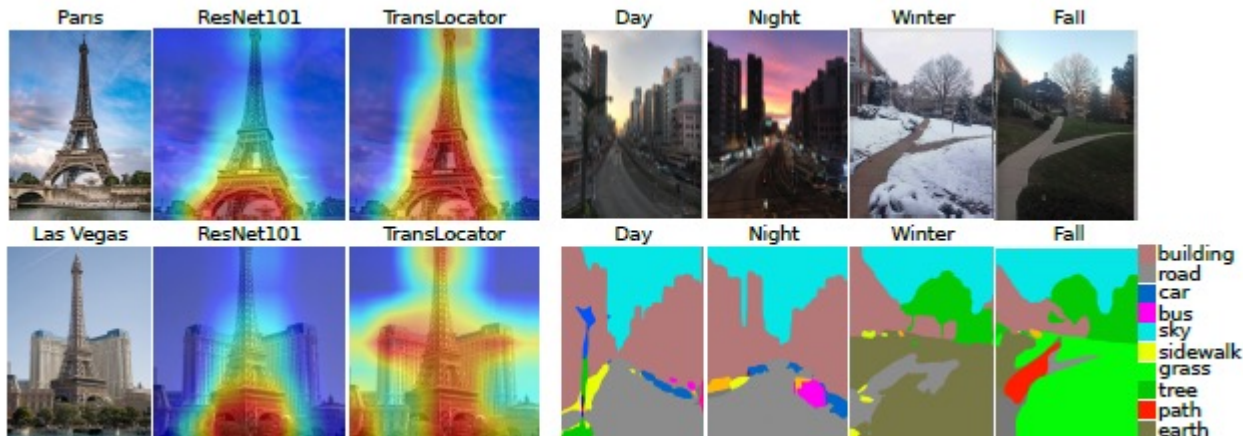
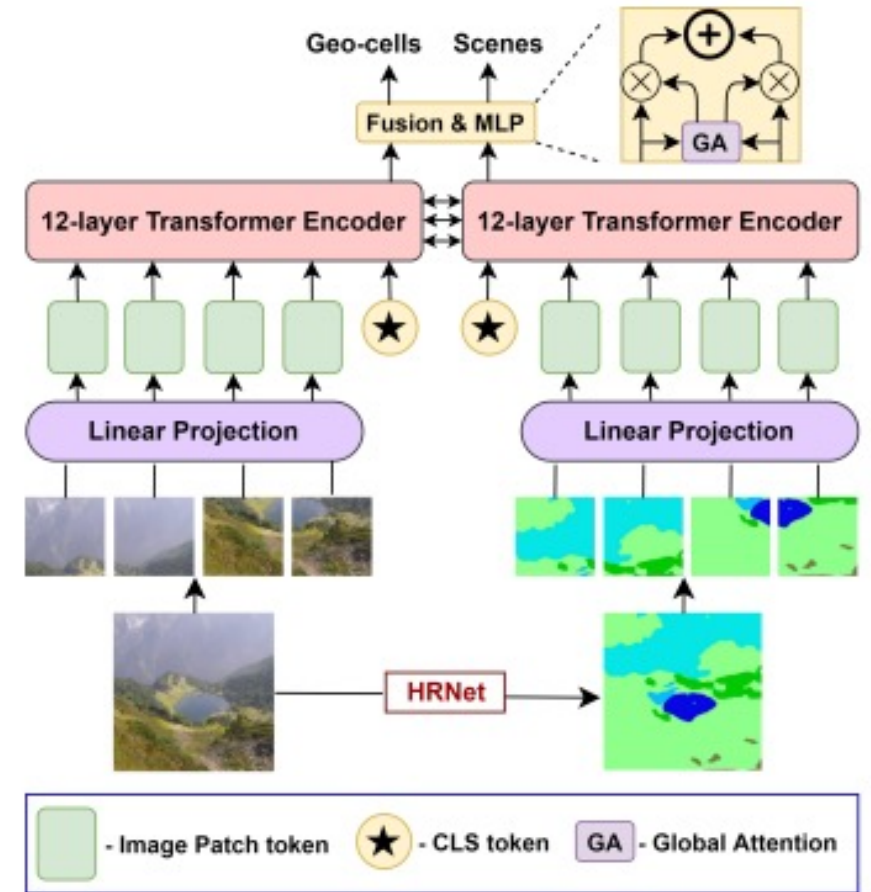
[1] Pramanick, Shraman, et al. "Where in the World is this Image? Transformer-based Geo-localization in the Wild." *ECCV*, 2022.

[2] Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020.

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. 2019

Representative Classification-Based Work: Translocator [1]

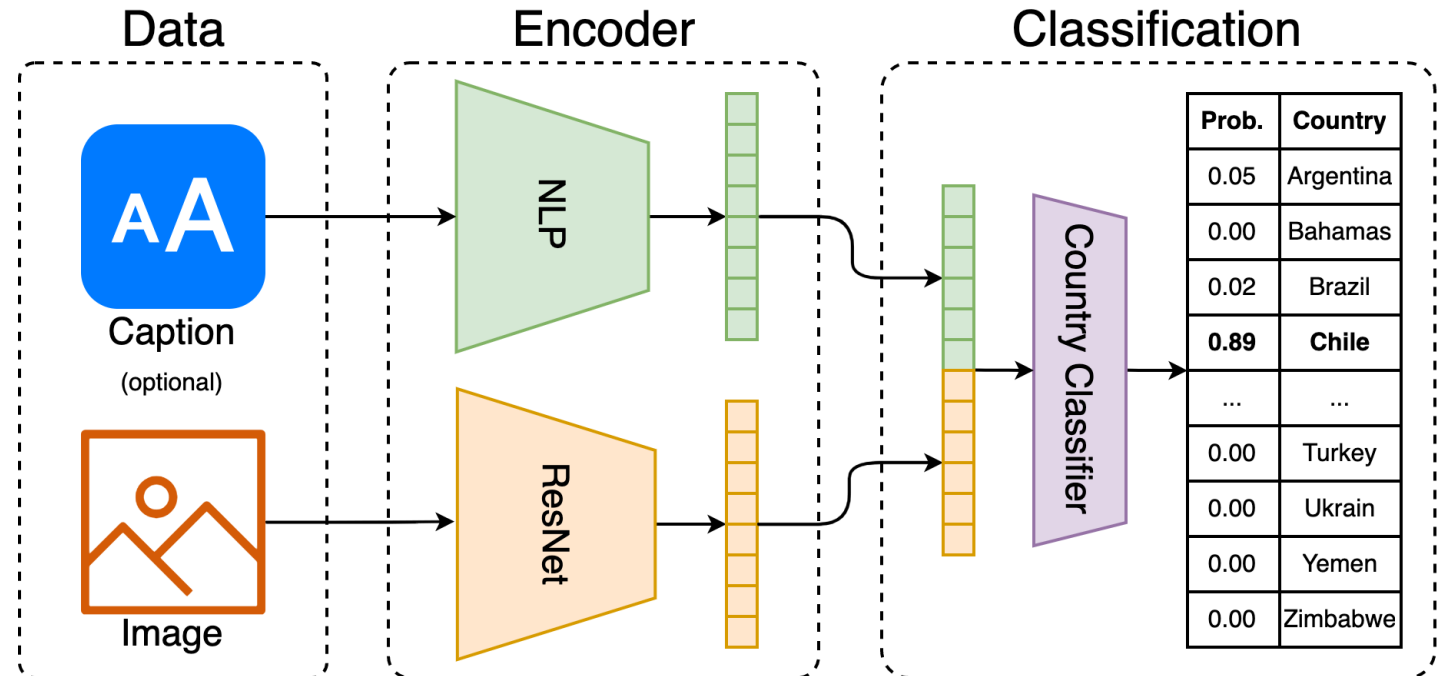
- Contain two parallel transformer branches, one for the RGB image and the other for the corresponding semantic map.
 - Sum the CLS tokens of each branch after every transformer encoder layer. At each layer, the CLS token is considered as an abstract global feature representation. This strategy is as effective as concatenating all feature tokens, but avoids quadratic complexity. Once the CLS tokens are fused, the information will be passed back to patch tokens at the later transformer encoder layers.
- Multi-Modal Fusion: The attention module has two major parts: modality attention generation and weighted concatenation.



[1] Pramanick, Shraman, et al. "Where in the World is this Image? Transformer-based Geo-localization in the Wild." *ECCV*, 2022.

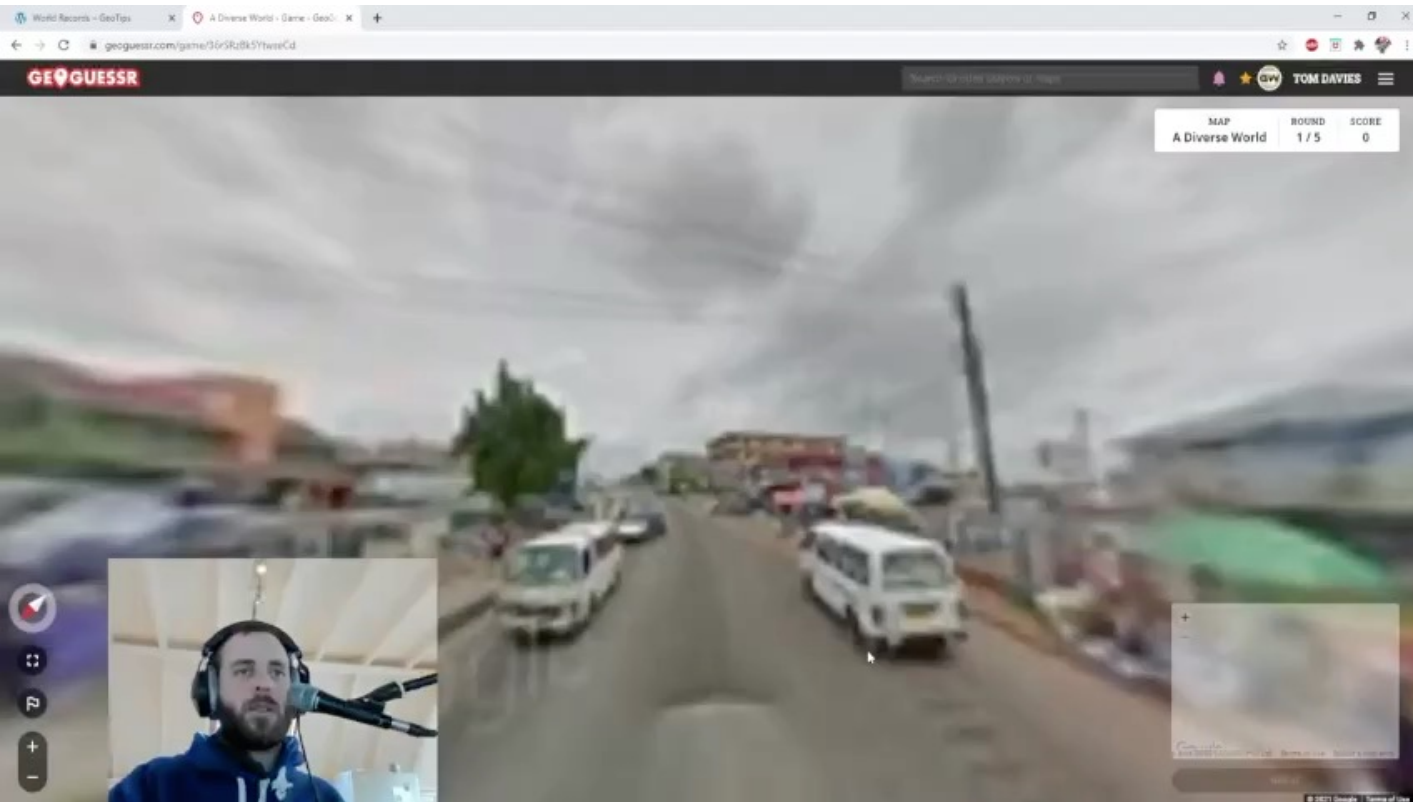
#3: Multi-Modal Geo-Localization

- In addition to images, text labels can be available information to aid geo-localization.
- The goal is to leverage available query data modalities (such as image and text) to improve geo-estimation accuracy.



Representative Work: G3 (Geolocation via Guidebook Grounding) [1]

GeoGuessr: given a random Google Street View panorama, guess where it is in the world.



- [GeoGuessr, Explore The World!](#)
- [GeoGuessr – The Top Tips, Tricks and Techniques](#)
- [Geolocation of Infrastructure Destruction in Cameroon: A Case Study of Kumbo and Kumfutu](#)
- [1] Luo, Grace, et al. "G³: Geolocation via Guidebook Grounding." *arXiv preprint arXiv:2211.15521* (2022).



Spanish bollards are fairly unique. They feature the standard European black and white bollard with a **bright yellow**, narrow rectangle encased in the black section of the bollard.



Portugal uses these fairly generic bollards. The front contains a vertical, **white stripe** encased in the black section.

- GeoGuessr is a popular geolocation game where the user is placed into a navigable StreetView scene and must predict either the GPS coordinates or the country (depending on the gameplay mode).
- The multimodal dataset, which includes images and texts, are extracted from a popular guide for playing the GeoGuessr game - "[GeoGuessr – The Top Tips, Tricks and Techniques](#)."

Representative Work: G3 (Geolocation via Guidebook Grounding) [1]

GT Location: Thailand



Guidebook Clues

Always be sure to look at both sides of **Thai** road markers and remember that **Thai** drivers drive on the left when they see the marker information.

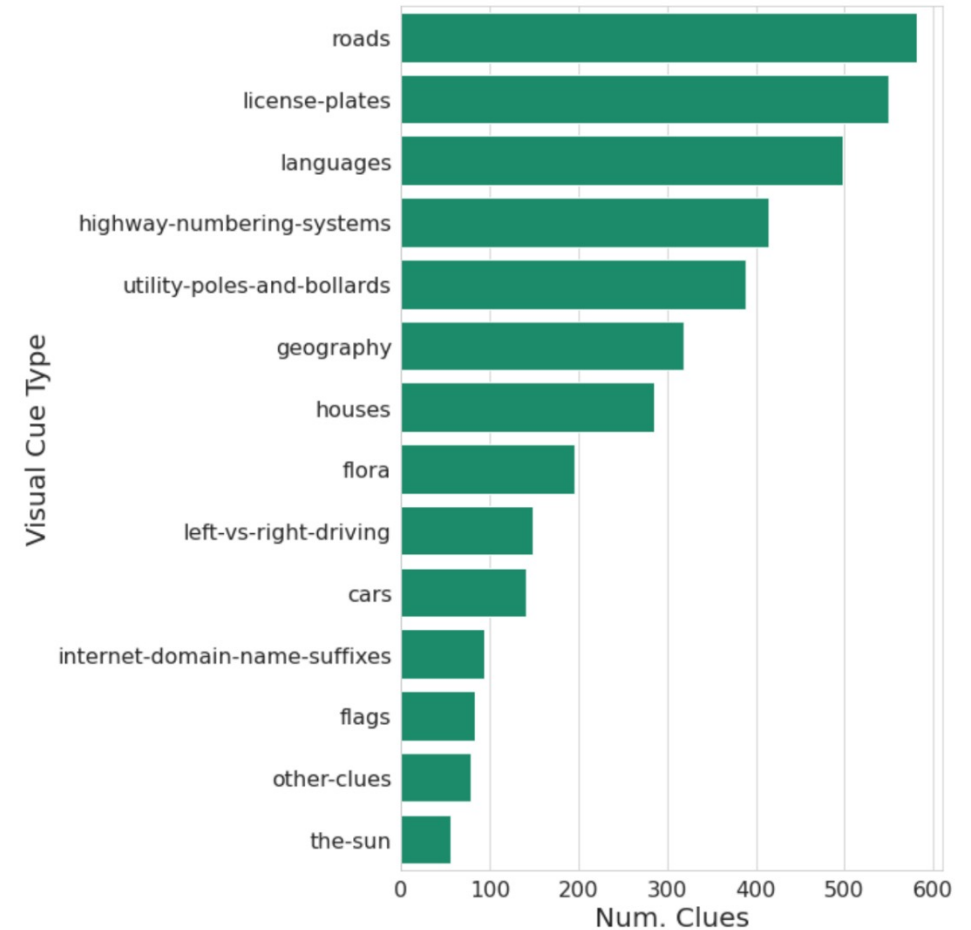
Most **Thai** roads will have some form of yellow central line as well as continuous white edge lines.

Many **Thai** houses can also look quite affluent and be two-storey and fairly large.

This contrasts the very small circles seen on the end of most **Thai** characters.

Posts that hold all types of **Thai** signs are unique in the world as they are wooden, painted white and importantly they have a section painted black on their base.

An example from the multimodal dataset of StreetView Images and relevant Guidebook Text. Note how some clues are grounded in the image. The image from Thailand depicts cars driven on the left side, roads with yellow center lines, and two-story houses.



Number of clues associated with each visual cue type

- [GeoGuessr – The Top Tips, Tricks and Techniques](#)

[1] Luo, Grace, et al. "G³: Geolocation via Guidebook Grounding." *arXiv preprint arXiv:2211.15521* (2022).

Representative Work: G3 (Geolocation via Guidebook Grounding) [1]

Multimodal embedding for classification:

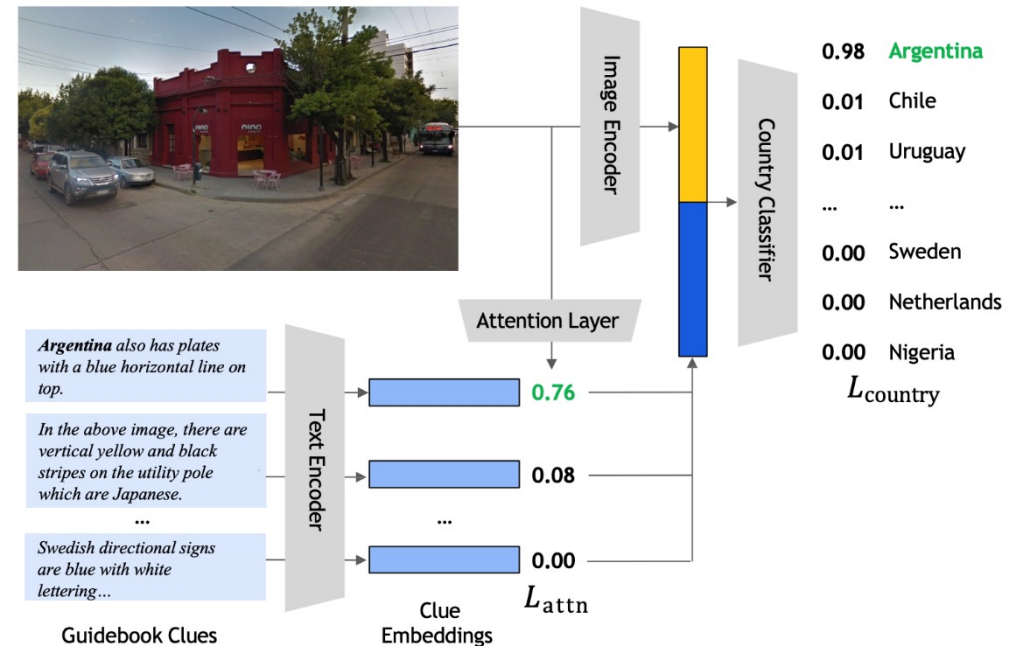
- Image embedding is obtained from a frozen CLIP RN50x16 encoder
- Clue embedding is computed by applying a frozen RoBERTa model to clues
- Passing the image embedding to an attention block yields attention score for each clue:

$$f_{attn}(d) = \text{ReLU}(W \cdot f_{CLIP}(d) + b)$$

- Take the weighted average of clue embeddings using the attention scores gives image-specific textual clue representation:

$$\hat{G} = \frac{1}{|G|} \sum_{i=1}^{|G|} \sigma(f_{attn_i}(d)) \times G_i$$

- The image embedding and clue representation are concatenated and passed to a classification block



Overview of G³ approach.

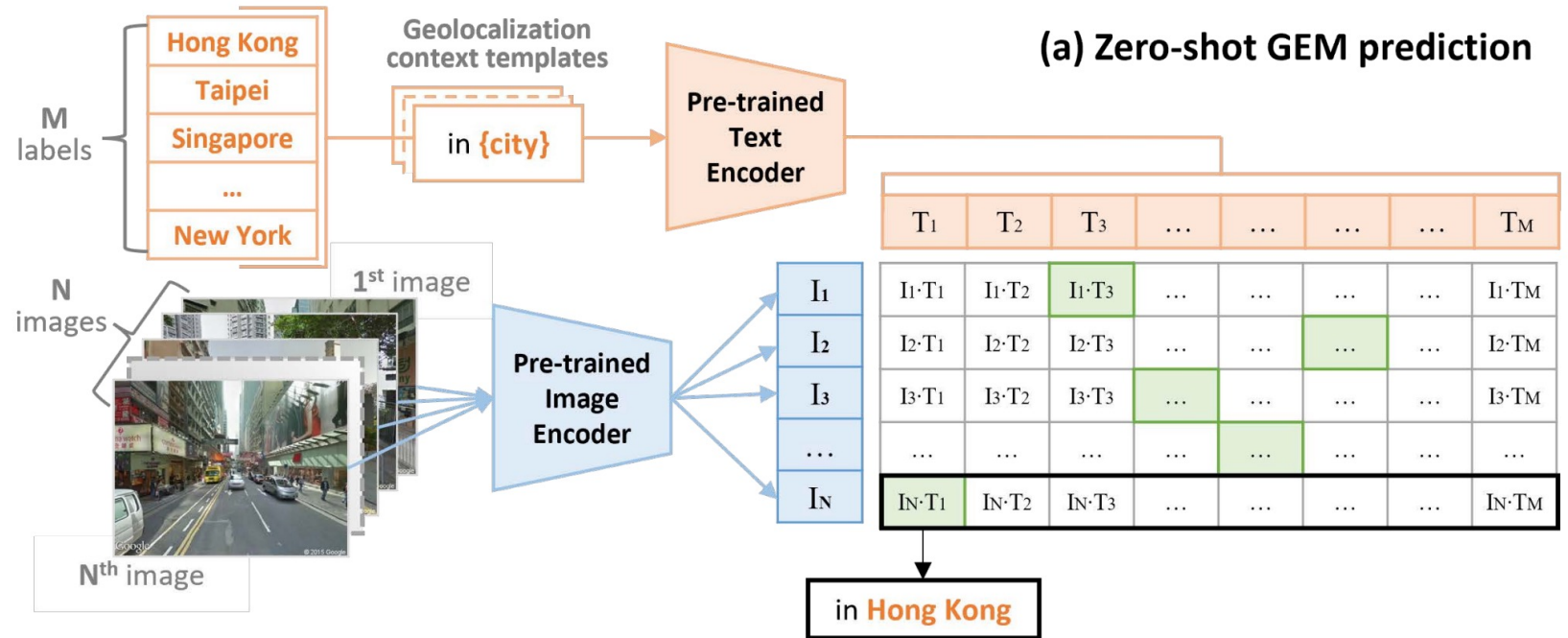
Model	Top-1	Top-5	Top-10
CLIP Nearest Neighbor	0.4336	0.6858	0.7806
CLIP Linear Probe	0.6081 ± 0.001	0.8789 ± 0.003	0.9417 ± 0.001
ISN	0.6527 ± 0.015	0.8817 ± 0.004	0.9379 ± 0.004
G ³ (Ours)	0.7031 ± 0.002	0.9178 ± 0.004	0.9618 ± 0.002

Table: Country Classification Accuracy

Representative Work: IM2City [1]

IM2City shows that a street view image can be geo-localized at a city level in the global scale, based on multi-modal learning with natural language and computer vision.

The image encoder from the CLIP-ViT-L/14 model (without the final classification layer) is used as the visual feature extractor. Specifically, for each image, a feature vector is output by this pre-trained image encoder, and then fed into a linear classifier for model training.



Zero-shot geo-localization framework

Text encoding:

$$T_{[n,d_e]}^{output} = \left\| T_{[n,d_t]}^{input} \times W_{[d_t,d_e]}^T \right\|^2$$

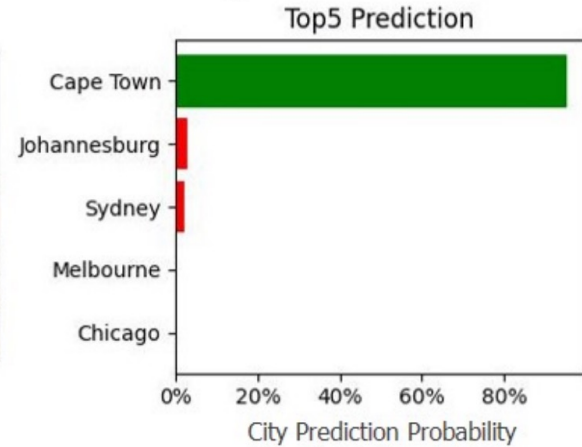
Image encoding:

$$I_{[n,d_e]}^{output} = \left\| I_{[n,d_i]}^{input} \times W_{[d_i,d_e]}^I \right\|^2$$

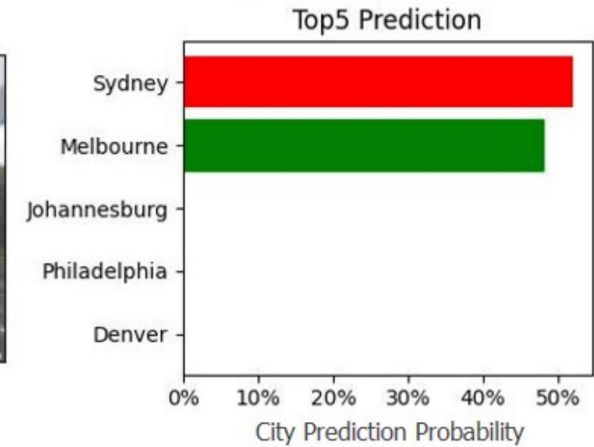
[1] Wu, Meiliu, and Qunying Huang. "IM2City: image geo-localization via multi-modal learning." *ACM SIGSPATIAL*. 2022.

Representative Work: IM2City [1]

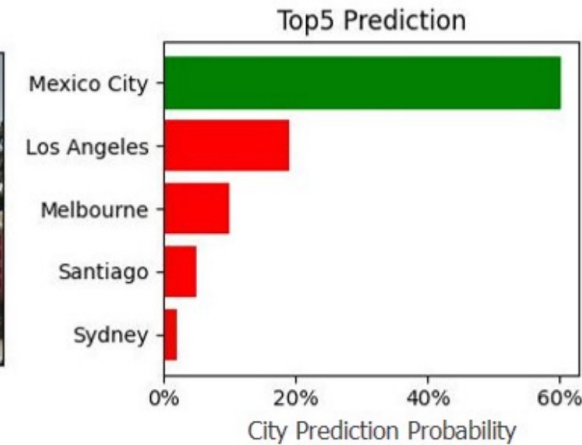
City: Cape Town Prediction: Cape Town



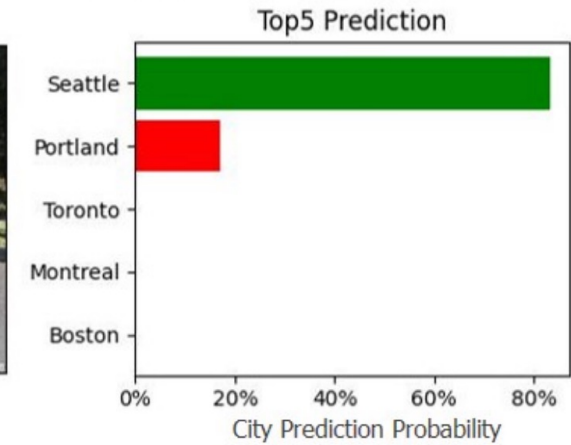
City: Melbourne Prediction: Sydney



City: Mexico City Prediction: Mexico City



City: Seattle Prediction: Seattle



[1] Wu, Meiliu, and Qunying Huang. "IM2City: image geo-localization via multi-modal learning." *ACM SIGSPATIAL*. 2022.

Outline

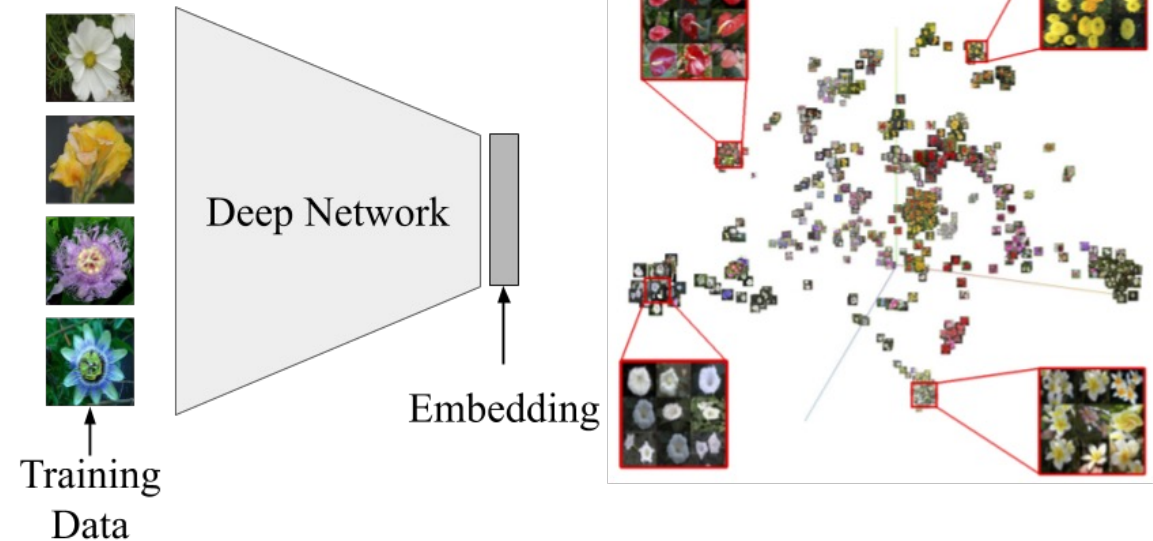
- Image-Based Geo-Localization
- Coarse Search: Approaches and Comparison
- Coarse Search: Trends
- **Fine Alignment**
- Use Cases
- Conclusion

Fine Alignment: Estimating camera pose (Location and Orientation)

Offline: Build image database and 3D Model/
Map for area of regard

Live: Estimate camera pose for a query image:

- **Image retrieval by search from database**
- Detect, describe, and match features to establish correspondences
- Estimate camera pose by resection using 2D-3D (image to model) correspondences



- Jégou, et al. "Aggregating local descriptors into a compact image representation." CVPR, 2010.
- Torii, et al. "24/7 place recognition by view synthesis." CVPR, 2015.
- Mithun, et al. "RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization." ACM Multimedia, 2020.

Cross-season / Cross-time Fine Alignment

Offline: Build image database and 3D Model/
Map for area of regard

Live: Estimate camera pose for a query image:

- Image retrieval by search from database
- **Detect, describe, and match features to establish correspondences**
- Estimate camera pose by resection using 2D-3D (image to model) correspondences



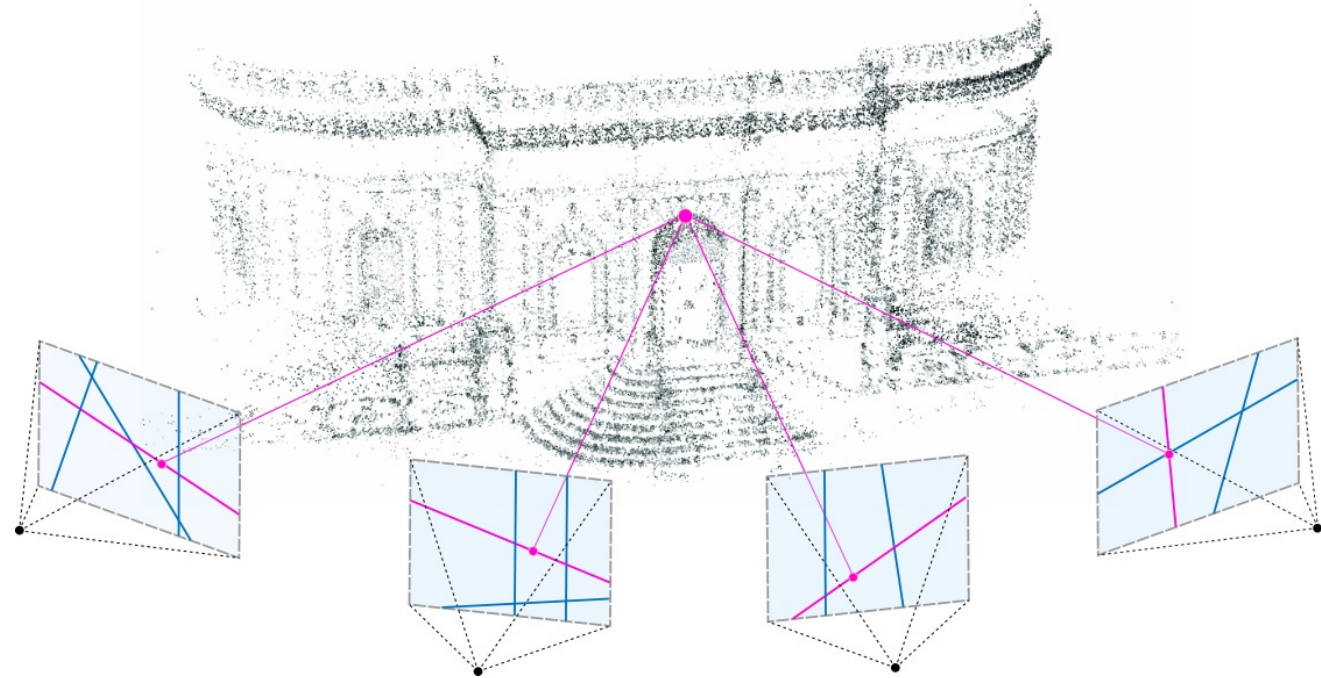
- Lowe. "Distinctive image features from scale-invariant keypoints." IJCV, 2004.
- Rublee, et al. "ORB: An efficient alternative to SIFT or SURF." ICCV, 2011.
- DeTone, et al. "Superpoint: Self-supervised interest point detection and description." CVPR, 2018.
- Sarlin, et al. "Superglue: Learning feature matching with graph neural networks." CVPR, 2020.

Cross-season / Cross-time Fine Alignment

Offline: Build image database and 3D Model/
Map for area of regard

Live: Estimate camera pose for a query image:

- Image retrieval by search from database
- Detect, describe, and match features to establish correspondences
- **Estimate camera pose by resection using 2D-3D (image to model) correspondences**

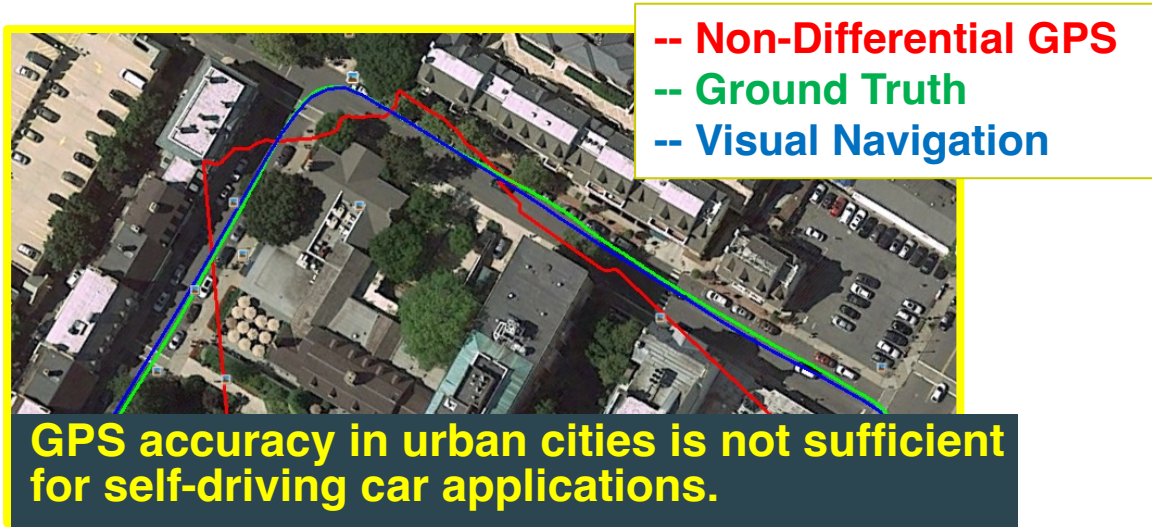


Outline

- Image-Based Geo-Localization
- Coarse Search: Approaches and Comparison
- Coarse Search: Trends
- Fine Alignment
- **Use Cases**
- Conclusion

Application: Navigation for Self-Driving Car

- **Challenge:** Automotive industry uses costly and bulky 3D LIDAR with GPS on each vehicle to geo-localize the vehicle within a 3D geo-referenced database, which is also constructed using LIDAR sensors.
- **Solution:** Using cameras instead of LIDARs for this geo-localization process can significantly reduce the sensor cost on each ground vehicle.
- **3D Geo-Referenced Database:**
 - Ground Camera Collection: Cross-Time
 - Aerial Camera Collection: Cross-View
 - LiDAR Collection: Cross-Modal



H. Chiu et al. "Night-Time GPS-Denied Navigation and Situational Understanding Using Vision-Enhanced Low-Light Imager", Joint Navigation Conference, 2023
V. Murali et al., "Utilizing Semantic Visual Landmarks for Precise Vehicle Navigation", IEEE ITSC 2017.
H. Chiu et al., "Sub-Meter Vehicle Navigation Using Efficient Pre-Mapped Visual Landmarks", IEEE ITSC 2016.

Application: Large-Scale Augmented Reality

- **Challenge:** Estimating highly-accurate 3D pose (position and orientation) of the user's view is necessary for AR, which offers simulated insertions mixed with a live video feed of user's real view.
 - GPS or traditional visual navigation (SLAM) algorithms cannot provide required localization accuracy requirement to large-scale AR.
- **Solution:** Visual geo-localization to a pre-built 3D geo-referenced database offers an appealing solution to this application.



Niluthpol Mithun, Kshitij Minhas, Han-Pang Chiu, Taragay Oskiper, Mikhail Sizintsev, Supun Samarasekera, Rakesh Kumar, "Cross-View Visual Geo-Localization for Outdoor Augmented Reality", VR, 2023.

Sizintsev, Mikhail, Niluthpol Chowdhury Mithun, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. "Long-Range Augmented Reality with Dynamic Occlusion Rendering." *ISMAR 2021*.

Han-Pang Chiu et al., "Augmented Reality Driving Using Semantic Geo-Registration", IEEE VR 2018.

Application: GPS-Denied Aerial Navigation

- **Challenge:** Provide accurate aerial navigation solutions under long-term GPS outage – navigation solution drifts quickly without GPS
- **Solution:** Incorporate absolute information through visual geo-registration from each input video frame to rendered image (from 3D referenced database) into navigation system.
- It provides accurate (3D RMS error < 10 meters) and consistent solutions on large-scale GPS-denied scenarios.



H. Chiu et al. “Optimized Simultaneous Aided Target detection and Imagery based Navigation in GPS-Denied Environments”, Joint Navigation Conference, 2022.

H. Chiu, A. Das, P. Miller, S. Samarasekera, R. Kumar, “Precise Vision-Aided Aerial Navigation”, IEEE IROS 2014.

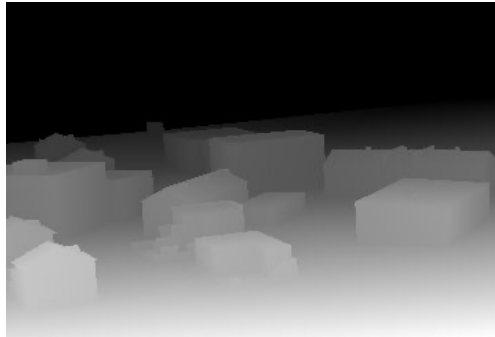
Geo-registration of video to site model ...

**Original
Video**

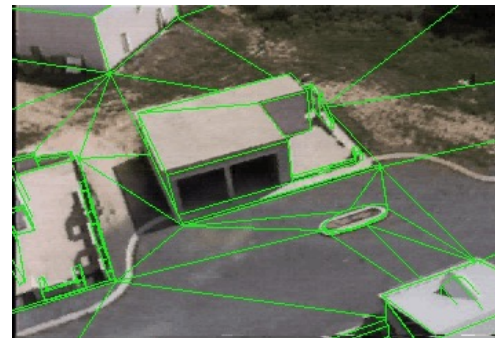
(video)



**Site
model**



**Geo-
registration
of video to
site model**



(video)



(video)

**Re-projection of video after merging with
model.**

Outline

- Image-Based Geo-Localization
- Coarse Search: Approaches and Comparison
- Coarse Search: Trends
- Fine Alignment
- Use Cases
- **Conclusion**

Agenda: Morning Talks

8:30 – 9:30am : Introduction of Generic Visual Geo-localization (Han-Pang Chiu, *in-person*).

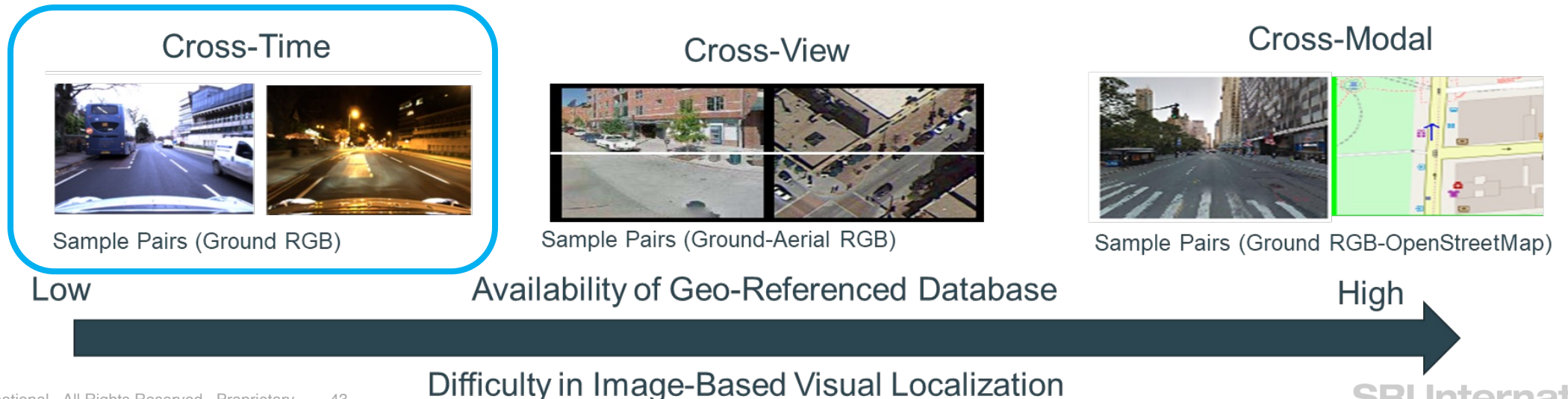
9:30 – 10:30 am : Recent Image Geo-localization Benchmark and Large-scale Real-world Scenarios (Carlo Masone /Gabriele Berton, *in-person*)

10:30 – 10:45: Coffee break

10:45 -11:15 am : R2Former: Unified Retrieval and Reranking Transformer for Place Recognition (Sijie Zhu, *in person*).

11:15 – 11:45 am : Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes (Mubarak Shah, *in person*).

11:45 – 12:30 pm: Lunch break



Agenda: Afternoon Talks

12:30 – 1:30 pm: Cross View and Cross-Modal Coarse Search and Fine alignment for Augmented Reality, Navigation and other applications (Rakesh Kumar, *in-person*).

1:30 – 2:30 pm: Toward Real-world Cross-view Geo-localization (Chen Chen/ Sijie Zhu, *in-person*)

2:30 – 3:00 pm: Vision-based Metric Cross-view Geo-localization (Florian Fervers, *in-person*).

3:00 – 3:30 pm: Coffee break

3:30 – 4:30 pm: Geometry-based Cross-view Geo-localization and Metric Localization for Vehicle (Yujiao Shi, *in-person*)

4:30 – 5:30 pm: Learning Disentangled Geometric Layout Correspondence for Cross-View Geo-localization (Waqas Sultani, *virtual*).

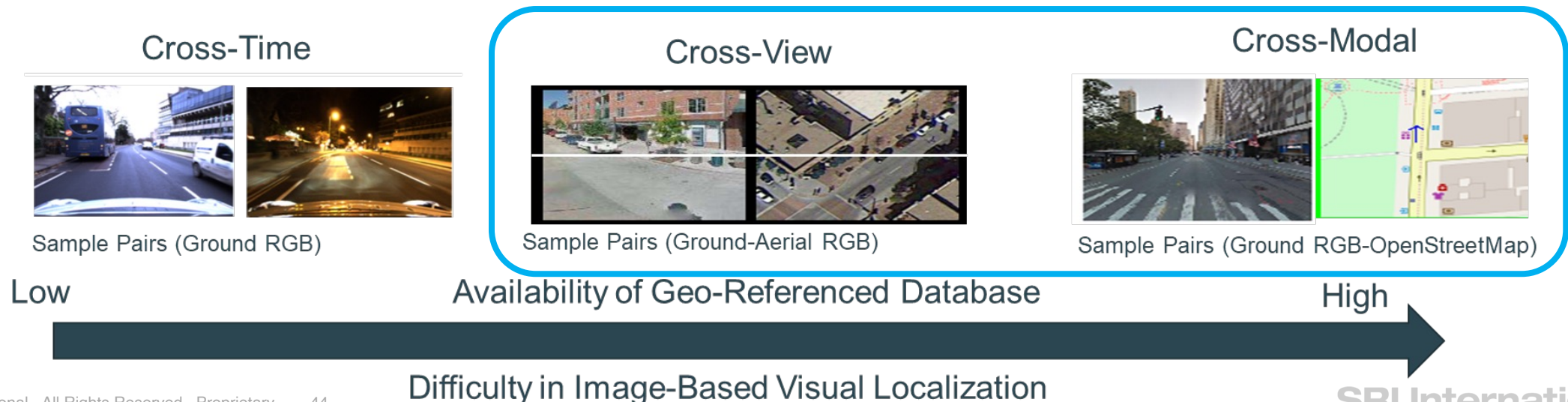


Image-Based Visual Geo-Localization

- Great recent progress toward real world-scale applications
- Huge potential and broad impact to many applications
- **Great research direction and topics for exploration!**



*This example is from SRI ONR WAR3D project

*Special thanks to my SRI colleagues, Angel Daruna and Tixiao Shan, for part of slide material

Supplementary Material

G3: Geolocation via Guidebook Grounding [1]

Street View Download 360

☰ **Street View Download 360** _ □ ×

📁 Location to save
C:\Files\Street View Panorama.png

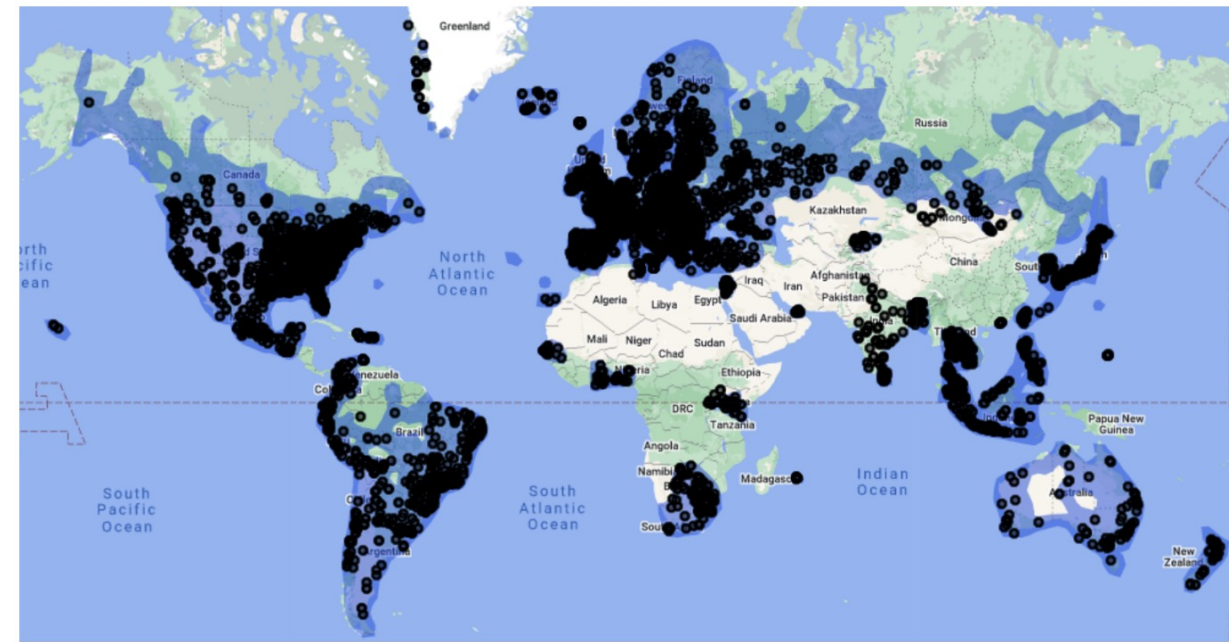
Single panorama Multiple panoramas

🔗 Panorama IDs or URLs
CAoSLEFGMVfpcE16SXixOUhHRU9xRIBSZI9EMEt4dGNrN3J3X3JMTFITNTJJMVhY
ST3dBfZEizgp-CgzZGWqfQ

Resolution
13312×6656

Download Panorama


Download completed 1/1



Street View Download 360: Multimodal Dataset details:

- Available for Windows, Mac OS and Linux.
- Resolution up to 16384×8192 px.
- Save as JPG, PNG, or WebP.
- Works for all types of Street View panoramas.
- Panorama: 88122
- Country: 90
- Clues: 3,182
- Cue types: 13
- Unique worlds: 3,712
- Clue average length: 14
- At least 426 panoramas for each country

Select panorama to preview
Street View Pano... >



📁 C:\Files\Street View Panorama...
File location

🔗 CAoSLEFGMVfpcE16SXixOUh...
Panorama ID

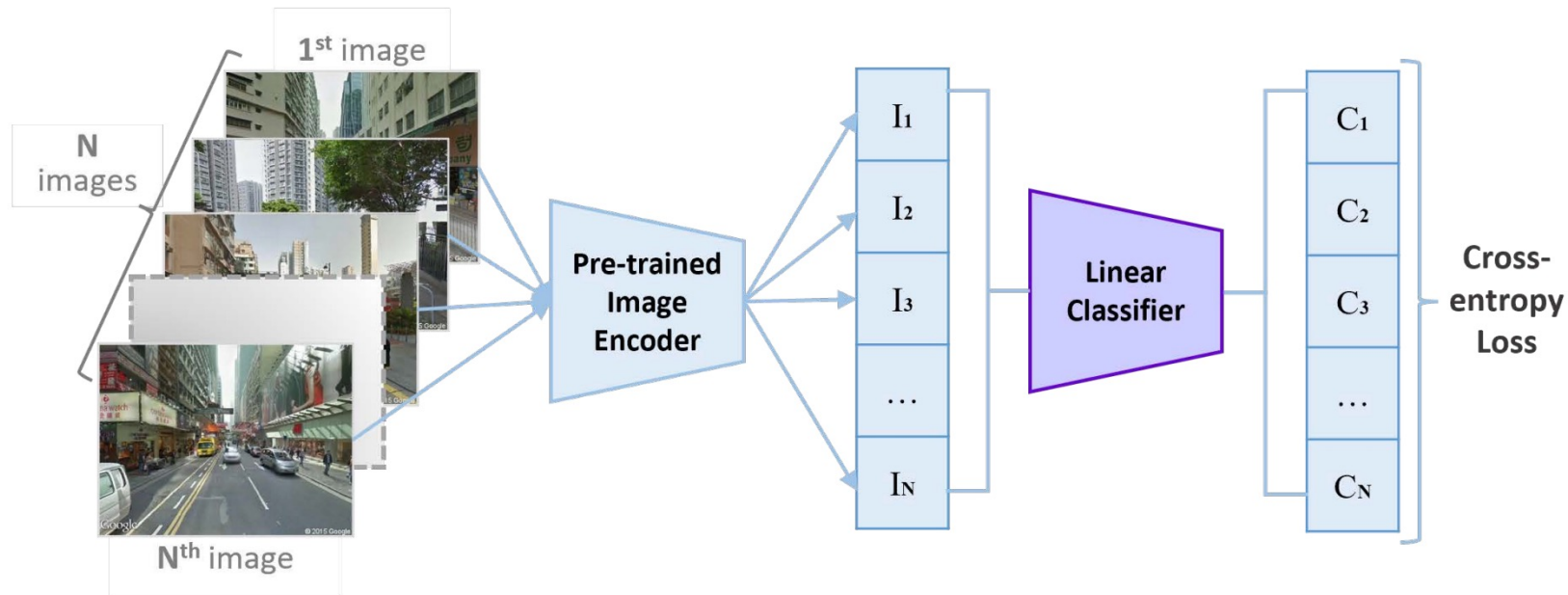
📍 40.275400, 9.592245
Geolocation

📅 2018-05-14
Date taken

📏 8000×4000
Resolution

IM2City: Image Geo-localization via Multi-modal Learning

(b) Linear Probe on GEM prediction



Linear-probing geo-localization framework

The image encoder from the CLIP-ViT-L/14 model (without the final classification layer) is used as the visual feature extractor. Specifically, for each image, a feature vector is output by this pre-trained image encoder, and then fed into a linear classifier for model training.