Cross View and Cross-Modal Coarse Search and Fine alignment for Augmented Reality, Navigation and other applications

Rakesh (Teddy) Kumar

Center for Vision Technologies
SRI International, Princeton, NJ, USA
rakesh.kumar@sri.com

June 18th, 2023

Cross-view matching talks in the afternoon

- **Cross-view and Cross-Modal Geo-localization:**
  a) 12.30 – 1.30 PM: Cross View and Cross-Modal Coarse Search and Fine alignment for Augmented Reality, Navigation and other applications, Rakesh (Teddy) Kumar, in-person.
  b) 1.30 PM – 2.30 PM: Toward Real-world Cross-view Geo-localization, Chen Chen/Sijie Zhu, in-person.
  c) 2.30 – 3.00 PM: Vision-based Metric Cross-view Geo-localization, Florian Fervers, in-person.

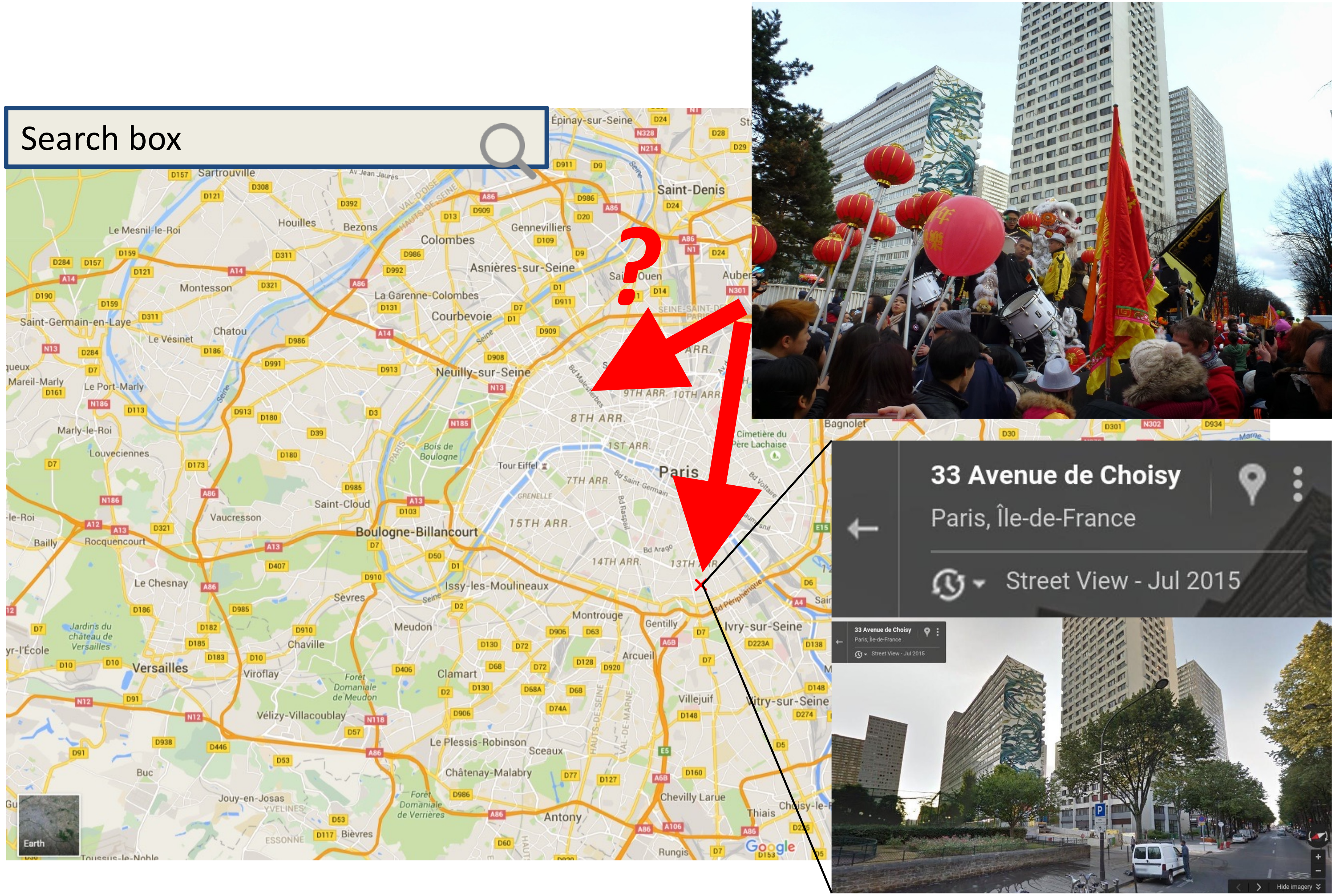
- **3:00 PM – 3:30 PM Coffee Break**

- **Cross-view and Cross-modal Geo-localization continuation:**
  a) 3.30 PM – 4.30 PM: Geometry-based Cross-view Geo-localization and Metric Localization for Vehicle, Yujiao Shi, in-person.
  b) 4.30 – 5.30 PM: Learning Disentangled Geometric Layout Correspondence for Cross-View Geo-localization Waqas Sultani, virtual.

Agenda

- Introduction to Problem
- Cross view matching approaches: Match 2D ground images to 2D overhead reference
  - Invariant features/ Project reference image to ground view
  - Project ground image to overhead view/ Bird's eye view
- Cross view matching: 2D ground images to 2.5 D overhead reference (2D reference image with 3D point cloud or terrain data)
- Cross modal matching: 2D ground images to LIDAR reference

Image based Geo-Localization



Missing hiker found based on photo he texted from Los Angeles area mountains

https://www.nbcnews.com/news/us-news/missing-hiker-found-based-photo-he-texted-los-angeles-area-n1264199?cid=sm_npd_nn_tw_ma

April 15th, 2021

Image based Geo-Location Applications

Visual Geo-localization

Autonomous Vehicles and Robotics

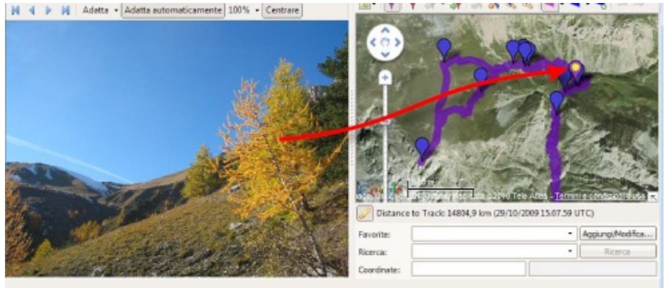
Robots

Appliances

Augmented Reality and Person Localization

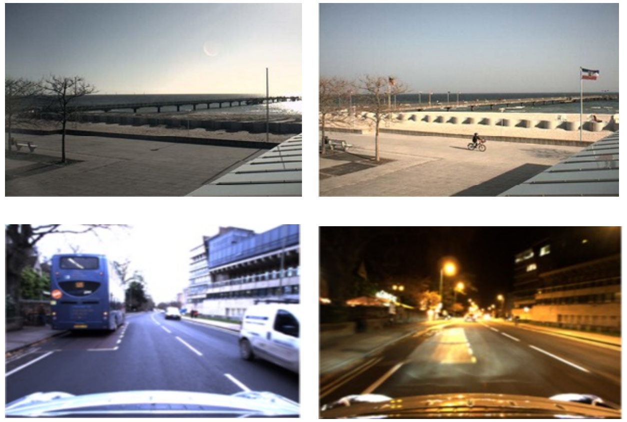
3D Modeling Outdoors & Indoors

Geo-tag images



Cross-Time, Cross-View, and Cross-Modal Matching of Ground Images to Reference Data

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

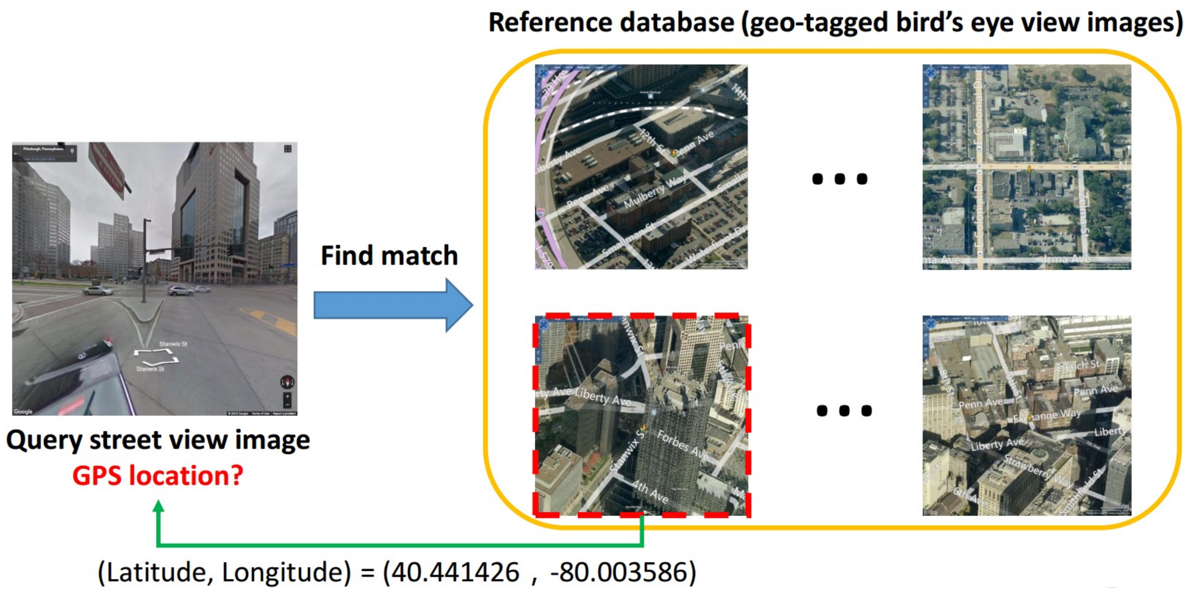
Difficulty in Image-Based Visual Localization based on Reference Data

Agenda

- Introduction to Problem
- Cross view matching approaches: Match 2D ground images to 2D overhead reference
  - Invariant features/ Project reference image to ground view
  - Project ground image to overhead view/ Bird's eye view
- Cross view matching: 2D ground images to 2.5 D overhead reference (2D reference image with 3D point cloud or terrain data)
- Cross modal matching: 2D ground images to LIDAR reference

Cross-View Geo-Localization

- **Problem:** Estimating the position and/or orientation of a camera at ground level given a largescale database of geo-tagged aerial (e.g., satellite) images.
- **Challenges:**
  - Extreme viewpoint change between ground and aerial images, traditional methods fail
  - Limited FOV in case using ground image from standard camera

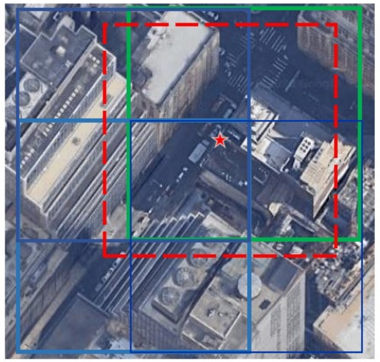


Benchmark Datasets for Cross-view Image Search

- **CVUSA:**
  - 44,416 pairs of ground and aerial images, Covers multiple cities across US
  - Ground Images: 360° panorama, 1232 x 224, Aerial Images: 750 x 750
  - GPS Tagged, Both ground and aerial images are north aligned
  - Reference image locations are at same location as of ground images
  - Workman S, Souvenir R, Jacobs N. 2015. Wide-Area Image Geolocalization with Aerial Reference Imagery. In: IEEE International Conference on Computer Vision (ICCV). 1–9. DOI: 10.1109/ICCV.2015.451.

- **CVACT:**
  - 128,334 pairs of ground and aerial images, Canberra (Australia)
  - Ground Images: 360° panorama, 1664 x 832, Aerial Images: 1200 x 1200, GPS Tagged, Both ground and aerial images are north aligned
  - Reference image locations are at same location as of ground images
  - Liu Liu, Hongdong Li et.al., Lending Orientation to Neural Networks for Cross-view Geo-localization, CVPR 2019 .

- **VIGOR:**
  - 90,618 aerial images (640 x 640), 238,696 ground panoramas (2048 x 1024),
  - Both ground and aerial images are north aligned
  - 4 references covering each query, raw GPS locations for ground images
  - S. Zhu, Taojiannan Yang, Chen Chen VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval, CVPR 2021.



Benchmark Datasets for Autonomous Driving Applications

Table 4. Datasets used to evaluate the proposed CVGL approach. Each ground data-frame consists of the vehicle's pose, camera images as well as intrinsic and extrinsic parameters. Data-frames are divided into disjoint cells with size 100m $\times$ 100m to measure aerial coverage. SD: Average scene duration in seconds.

Dataset	Region	Year	Scenes	Frames ($\times 10^3$)	SD (sec)	Cams	Cells	Orthophoto providers
Argoverse V1 [11]	Miami	$\leq$ 2019	53	12	22	9	71	Google Maps [3], Bing Maps [1]
	Pittsburgh	$\leq$ 2019	60	10	17	9	55	Google Maps [3], Bing Maps [1]
Argoverse V2 [45]	Austin	$\leq$ 2021	111	48	43	7	296	Google Maps [3], Bing Maps [1], Stratmap [5]
	Detroit	$\leq$ 2021	256	91	36	7	569	Google Maps [3], Bing Maps [1]
	Miami	$\leq$ 2021	703	245	34	7	811	Google Maps [3], Bing Maps [1]
	Palo Alto	$\leq$ 2021	43	136	34	7	157	Google Maps [3], Bing Maps [1]
	Pittsburgh	$\leq$ 2021	668	228	34	7	557	Google Maps [3], Bing Maps [1]
	Washington	$\leq$ 2021	262	90	34	7	553	Google Maps [3], Bing Maps [1], DCGIS [2]
Ford AV [6]	Detroit	2017	18	136	811	6-7	983	Google Maps [3], Bing Maps [1]
KITTI-360 [21]	Karlsruhe	2013	9	76	877	3	609	Google Maps [3], Bing Maps [1]
Lyft L5 [18]	Palo Alto	2019	398	50	25	6	88	Google Maps [3], Bing Maps [1]
Nuscenes [9]	Boston	2018	467	19	20	6	174	Google Maps [3], Bing Maps [1], MassGIS [4]
Pandaset [49]	Palo Alto	2019	35	3	8	6	87	Google Maps [3], Bing Maps [1]
	San Francisco	2019	65	5	8	6	93	Google Maps [3], Bing Maps [1]

From Fervers et.al., Uncertainty-aware Vision-based Metric Cross-view Geo-localization, CVPR 2023, https://arxiv.org/abs/2211.12145

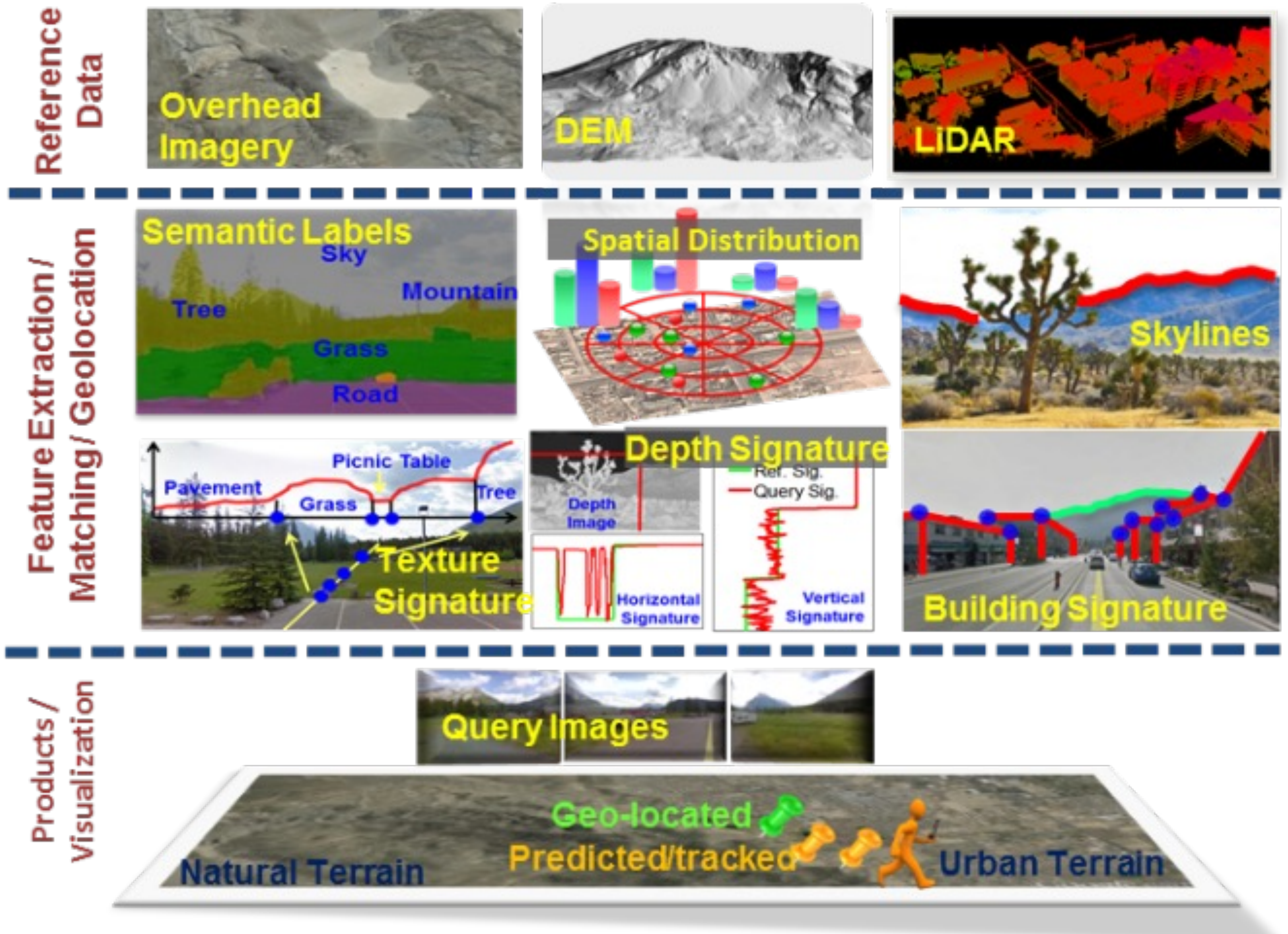
Geo-registration of ground imagery to overhead reference

Use Overhead Reference Data

Extract Terrain, 3D, Appearance, Semantic Features

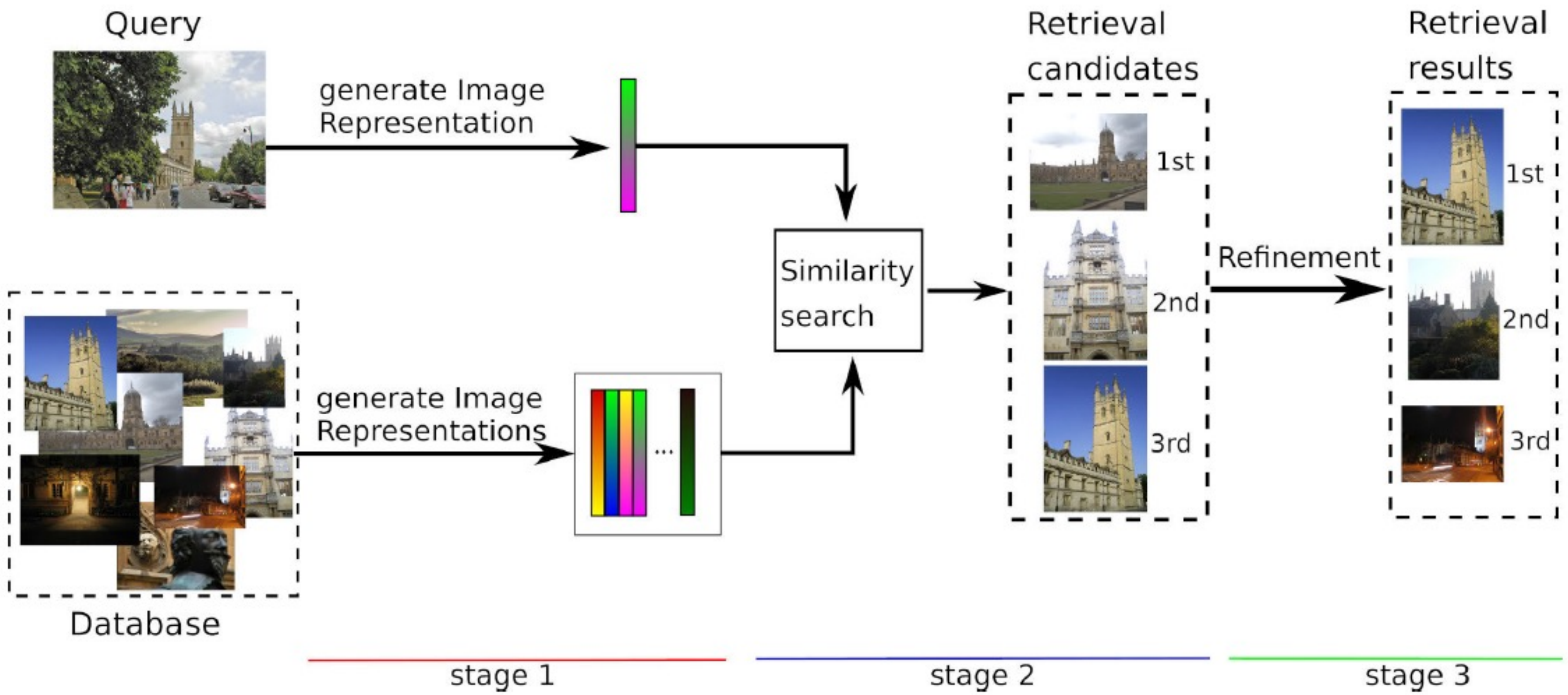
Match Features Extracted from Query Images

(Absolute Pose)



Estimate Absolute Orientation or Absolute Pose if looking at nearby objects

Feature Representation for Image Retrieval using Cross-view matching



Visual place recognition is commonly formulated as an image retrieval problem. The known places are collected in a database and a new image to be localized is called query. The place retrieval is performed in three logical stages.

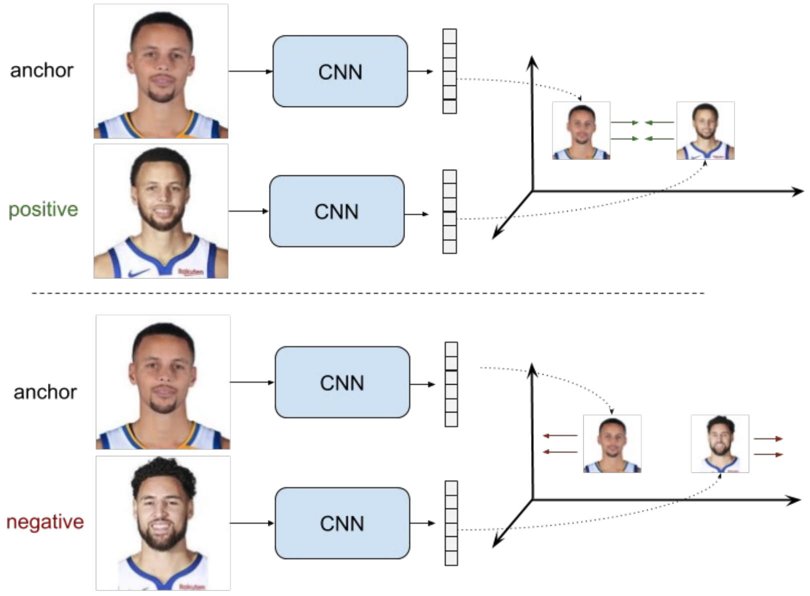
1) In the first stage, vector representations are generated for the query and the database images. From a practical perspective, the representation of the query is computed online, whereas the representations of the database images are computed offline.
2) The representation of the query is compared to those of the database images, to find the most similar ones (here only the top 3 are shown).
3) The best results of the comparison are further refined with post-processing techniques (here only the top3 are shown).

From: C. Masone and B. Caputo, "A Survey on Deep Visual Place Recognition," in IEEE Access, vol. 9, pp. 19516-19547, 2021, doi: 10.1109/ACCESS.2021.3054937.

Ranking Loss functions for Training Neural Network for Image Retrieval

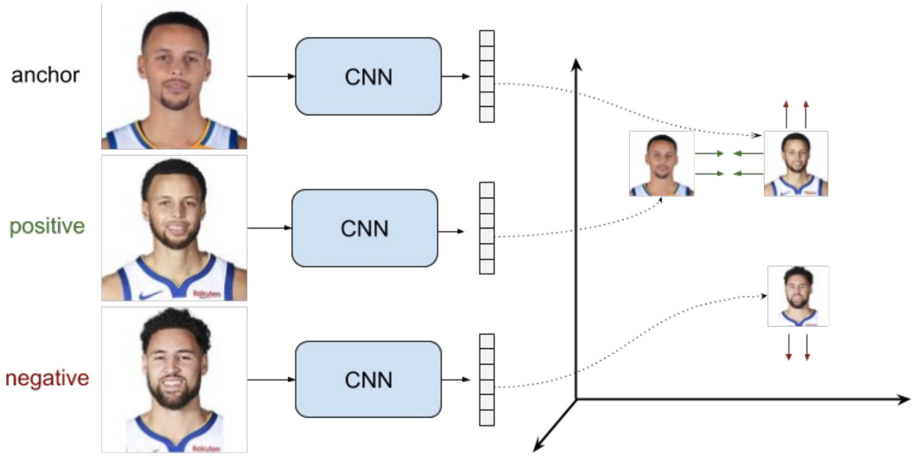
Pairwise Ranking Loss using Siamese Networks

$$L = \begin{cases} d(r_a, r_p) & if \quad PositivePair \\ max(0, m - d(r_a, r_n)) & if \quad NegativePair \end{cases}$$



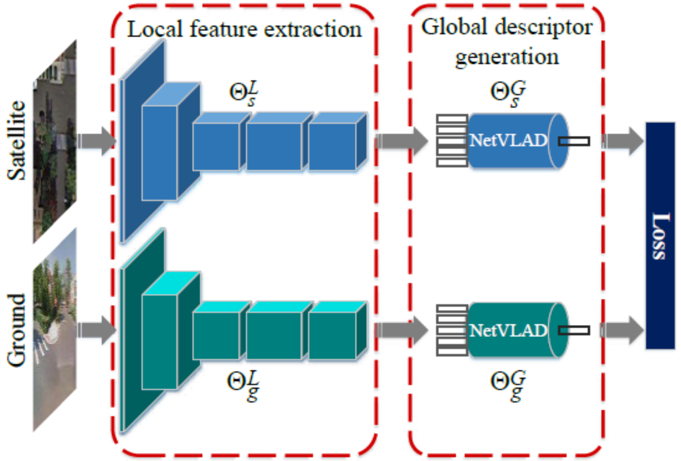
Triplet Ranking Loss

$$L(r_a, r_p, r_n) = max(0, m + d(r_a, r_p) - d(r_a, r_n))$$

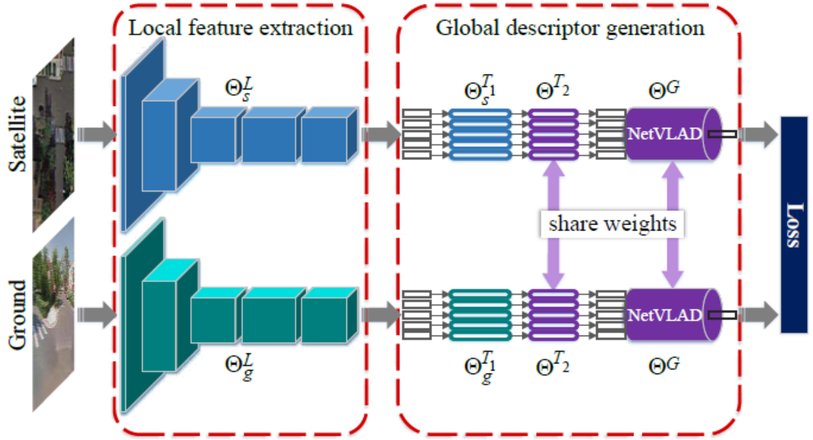


Using CNN's for Cross-view matching

- Siamese like architecture, containing two network branches of same architecture to operate on aerial and ground image
- Fully Convolutional Network to extract local feature vectors
- NetVLAD layer to extract global descriptors that are invariant to large viewpoint changes
  - NetVLAD, is a generalized trainable VLAD layer using neural networks, inspired by the "Vector of Locally Aggregated Descriptors" image representation.



CVM-Net-I

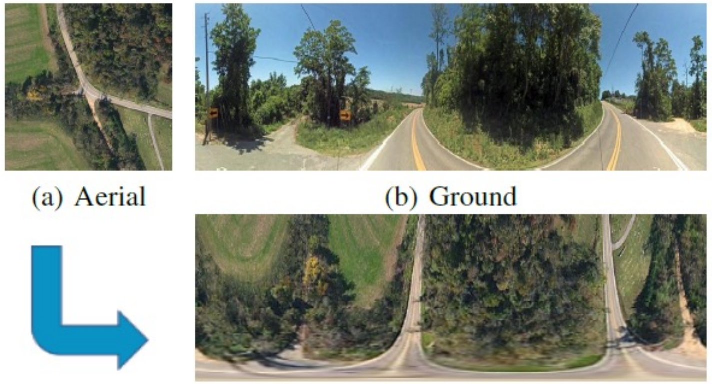


CVM-Net-II

Sixing Hu, et.al., CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization [CVPR 2018]
NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, [CVPR 2016]

Polar transforms and Spatial-Aware Feature Aggregation (SAFA) for Cross-View Image based Geo-Localization

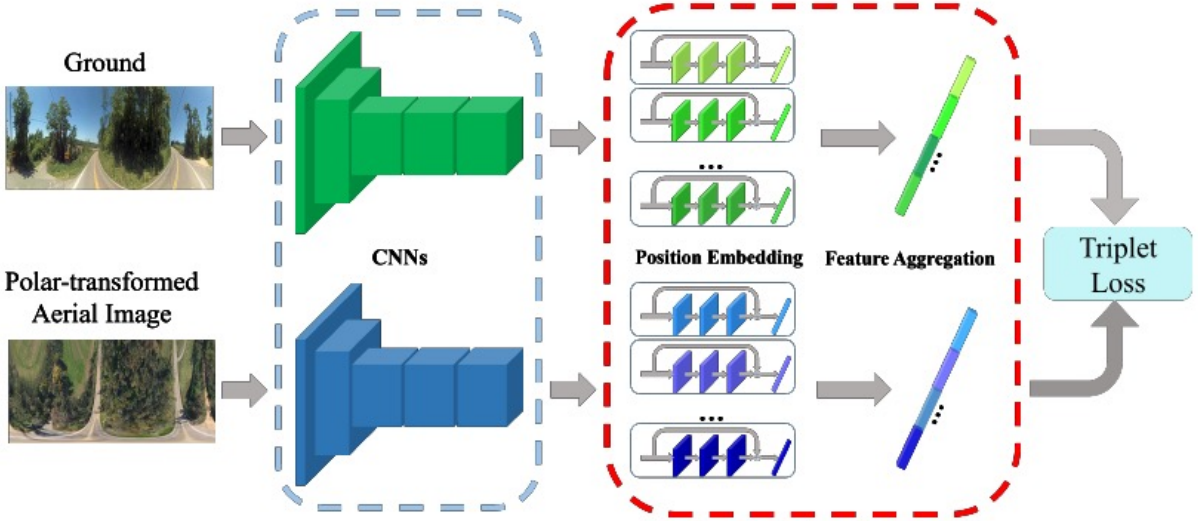
- **This paper provided a breakthrough in this field in terms of accuracies achieved**
- Basic Assumption:
  - The radial lines on an aerial image approximately correspond to the vertical lines on the matching ground image and,
  - The circular lines on the aerial image approximately correspond to the horizontal lines on the matching ground image
- Based on this assumption, a **Polar Transform** is applied on the aerial images



(a) Aerial (b) Ground

(d) Polar-transformed Aerial Image

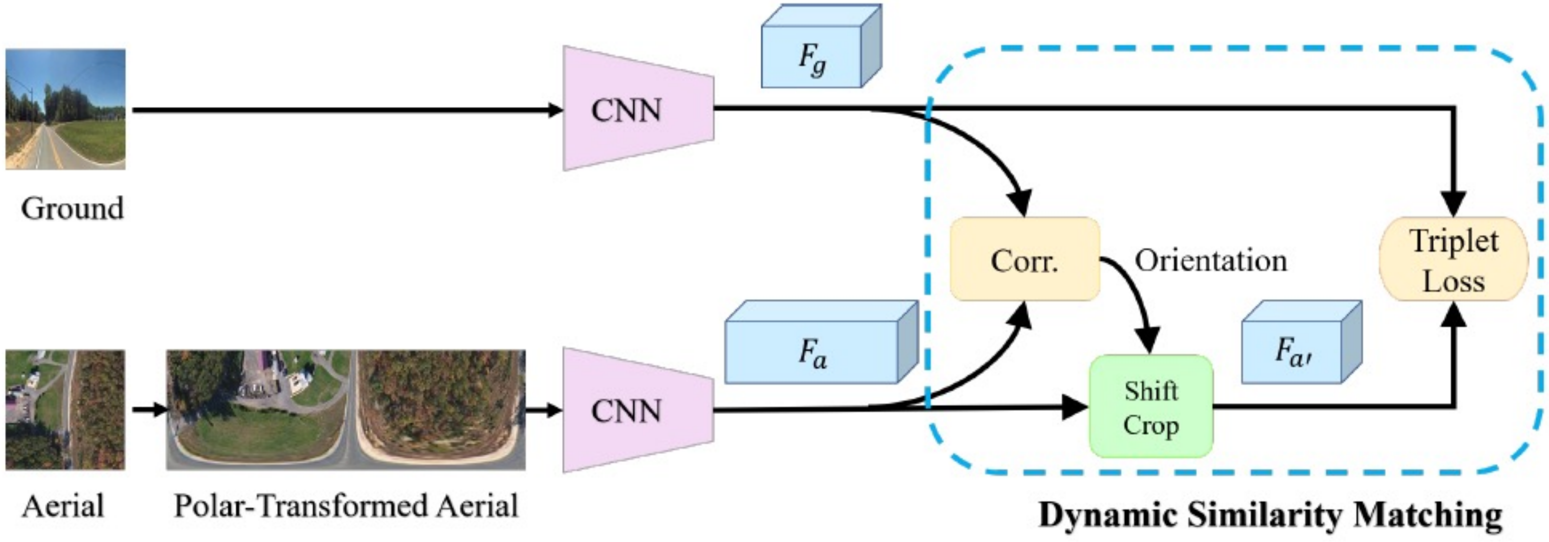
- As polar transformation doesn't take depth of the scene into account, it introduces distortion in the transformed images
- **Spatial-aware Position Embedding Module (SPE):**
  - Employs a max-pooling operator along feature channels to choose the most distinct object feature, and then adopts a Spatial-aware Importance Generator to generate a position embedding map.
  - Imposes an attention mechanism to select salient features while suppressing the features caused by the distortions
  - Multiple SPE modules are employed with the same architecture but different weights to generate multiple embedding maps to encode multiple input features



Y. Shi et.al. SAFA: Spatial-Aware Feature Aggregation for Cross-View Image based Geo-Localization [NeurIPS 2019]

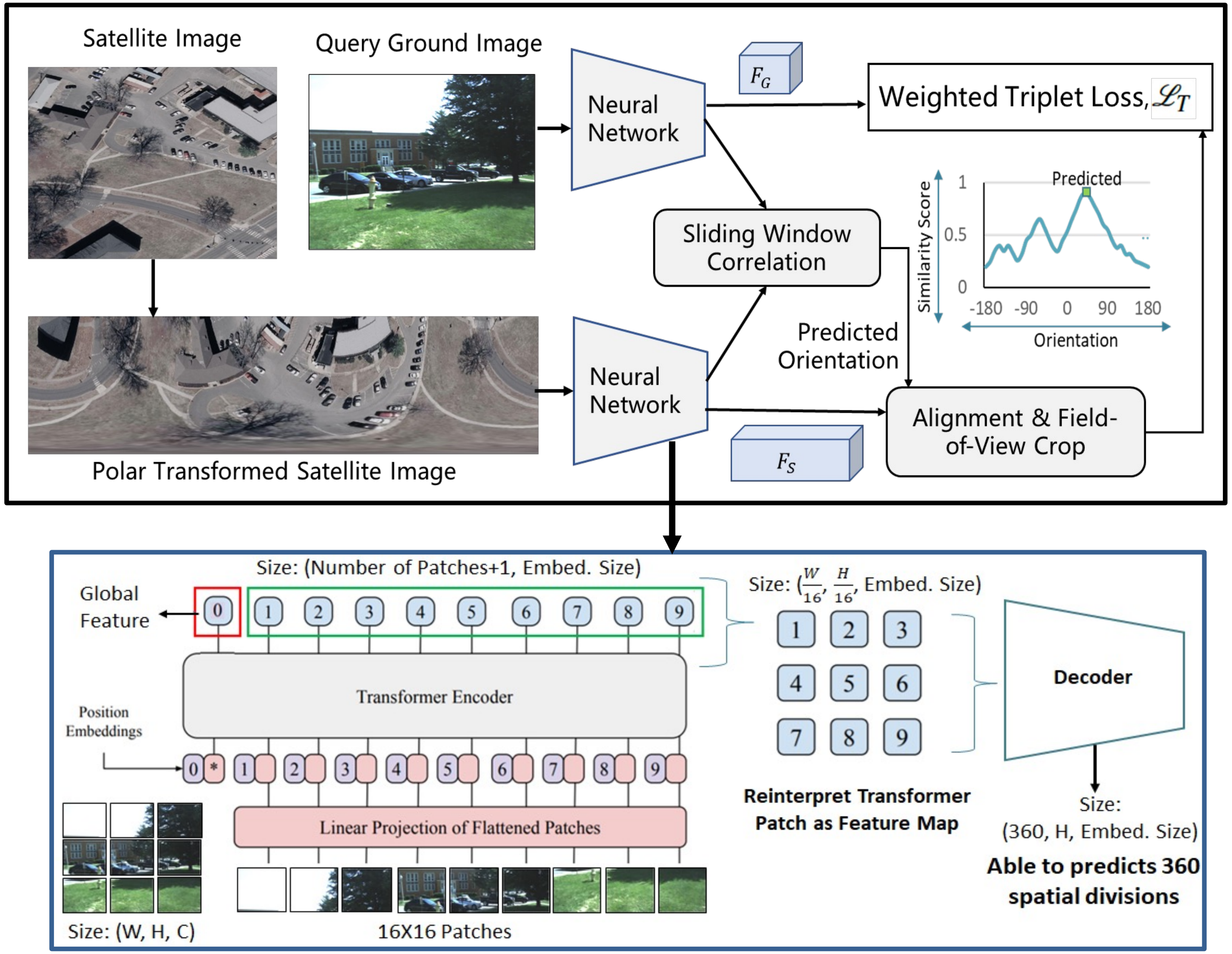
Joint Location and Orientation Estimation by Cross-View Matching

- Most of the approaches discussed till now assume that the orientation (azimuth angle) for the input ground image is known. (Even if not, orientation isn't estimated)
- First work to estimate both location and orientation via cross-view matching
- Also allow for query ground images with limited FOV
- They employ SAFA-like architecture combined with **Dynamic Similarity Matching (DSM)** Module to estimate cross-view orientation alignment.



Y. Shi e.al. Where am I looking at? Joint Location and Orientation Estimation by Cross-View Matching [CVPR 2020]

Geo-Registration using 2D Reference Data – Location and Orientation Estimation



(Transformer based) Neural Network Model

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Approach Overview

- The neural network model is trained using proposed orientation weighted triplet loss to simultaneously perform location and orientation estimation.
- Convolutional Neural Network (CNN) or Vision Transformer (ViT) Neural Network used as base model
- A decoder followed by ViT Encoder is used to increase the feature map spatial size to perform fine-grained orientation orientation.
- Street view images from search engine (e.g, Google, Bing) and corresponding aerial ref. ortho images are used for training.

Orientation weighted Triplet Loss: $\mathscr{L}_T = \mathscr{W}_{Ori} * \mathscr{L}_{GS}$

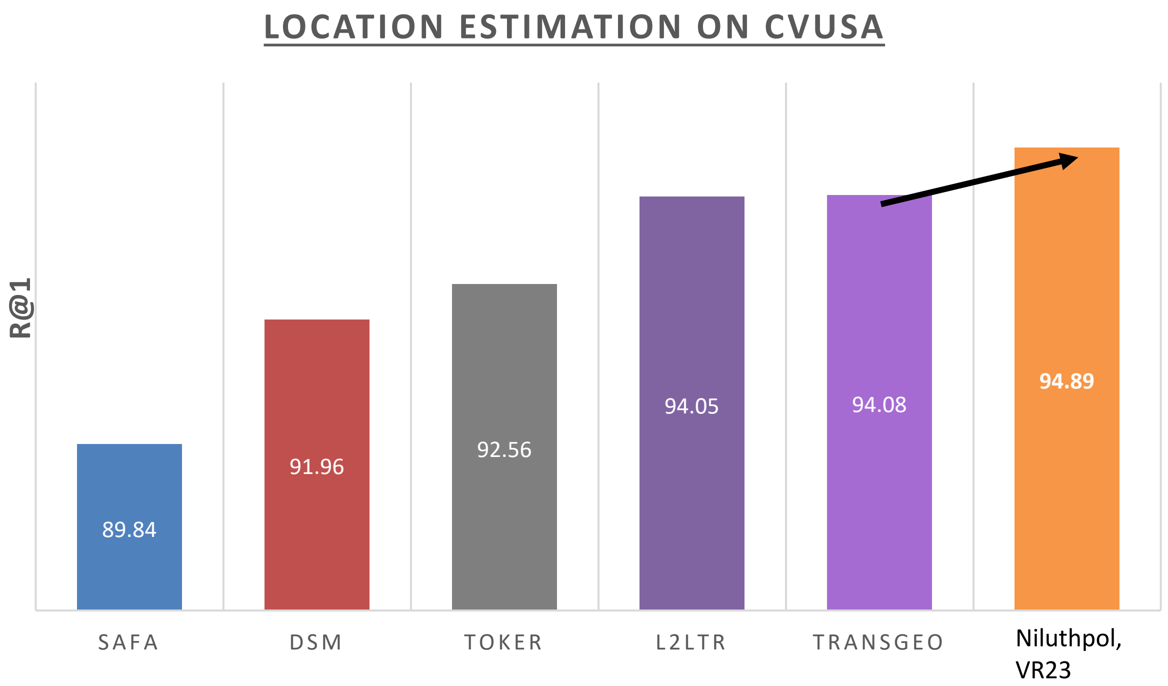
Triplet Loss: $\mathscr{L}_{GS} = \log\left(1 + e^{\alpha(||\mathbb{F}_G - \mathbb{F}_S||_F - ||\mathbb{F}_G - \mathbb{F}_{\hat{S}}||_F)}\right)$

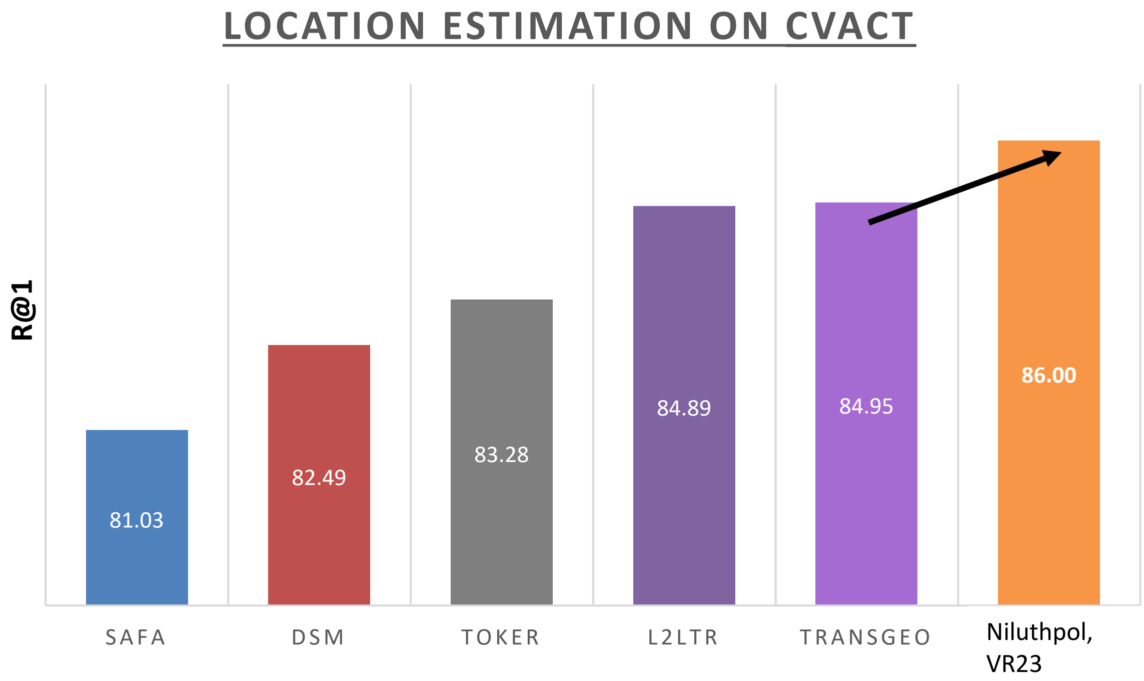
Orientation Weight Factor: $\mathscr{W}_{Ori} = 1 + \beta * \frac{\mathbb{S}_{Max} - \mathbb{S}_{GT}}{\mathbb{S}_{Max} - \mathbb{S}_{Min}}$

$\mathbb{F}_{\hat{S}}$ is a non-matching satellite image feature embedding for ground image feature embedding $\mathbb{F}_G$, and $\mathbb{F}_S$ is the matching satellite feature embedding.

$\mathbb{S}_{Max}$ and $\mathbb{S}_{Min}$ are the maximum and minimum value of similarity scores. $\mathbb{S}_{GT}$ is the similarity score at the ground-truth position.

Location Estimation Performance





- Achieves state-of-the-art performance in both CVUSA and CVACT datasets

[SAFA] Y. Shi, et al., “Spatial-aware feature aggregation for cross-view image based geo-localization” NeurIPS, 2019.
[DSM] Y. Shi, et al., “Where am I looking at? joint location and orientation estimation by cross-view matching”, CVPR 2020
[Toker] A. Toker, et al., “Coming down ´ to earth: Satellite-to-street view synthesis for geo-localization, CVPR 2021
[L2LTR] H. Yang, et. al., Cross-view geo-localization with layer-to-layer transformer, NeurIPS, 2021
[TransGeo] S. Zhu, et al., “TransGeo: Transformer is all you need for cross-view image geo-localization, CVPR 2022
[Niluthpol, VR23] M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Orientation Estimation on CVUSA

Comparisons of orientation estimation results with state-of-the-art methods and baselines on CVUSA.

- In row-3.1, We report performance for prediction range 64 (as reported in prior work DSM [28]) .
- In row-3.2, we present the performance of baselines implemented by us for prediction range 360. By default, Proposed (Full) is with Transformer based backbone. We also report with CNN backbone.

#	Method	Base Neural Network	Prediction Range	Orientation Error Range			
				2 Deg.	4 Deg.	6 Deg.	12 Deg.
3.1	L2LTR [36]	CNN	64	-	-	0.27	0.54
	DSM [28]	CNN	64	-	-	0.85	0.9
	Niluthpol VR23 (w/ L_T)	CNN	64	-	-	0.89	0.94
	Niluthpol VR23 (w/ L_T)	Transformer	64	-	-	0.94	0.98
3.2	DSM (Updated for 360)	CNN	360	0.88	0.93	0.93	0.95
	Niluthpol VR23 (w/o W_{ORI})	Transformer	360	0.77	0.93	0.97	0.98
	Niluthpol VR23 (w/ L_T)	CNN	360	0.89	0.95	0.96	0.97
	Niluthpol VR23 (w/ L_T)	Transformer	360	0.93	0.97	0.98	0.99

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

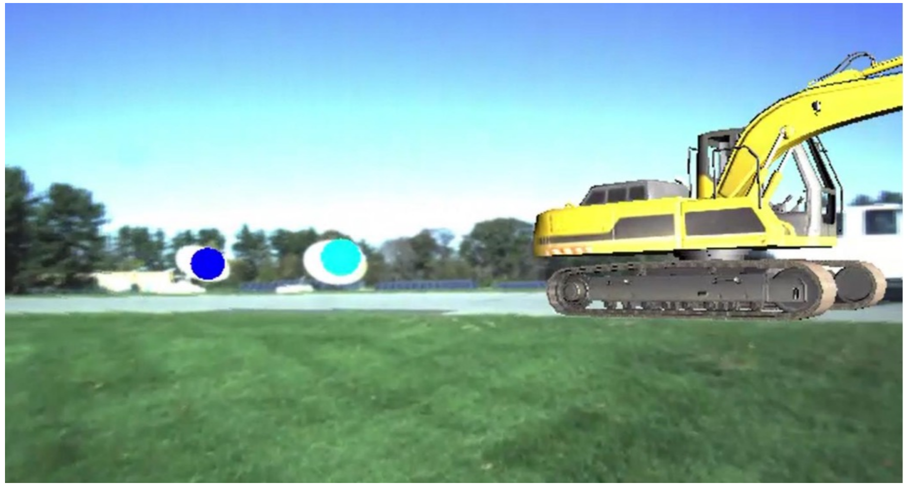
Cross-View Visual Geo-Localization for Outdoor Augmented Reality

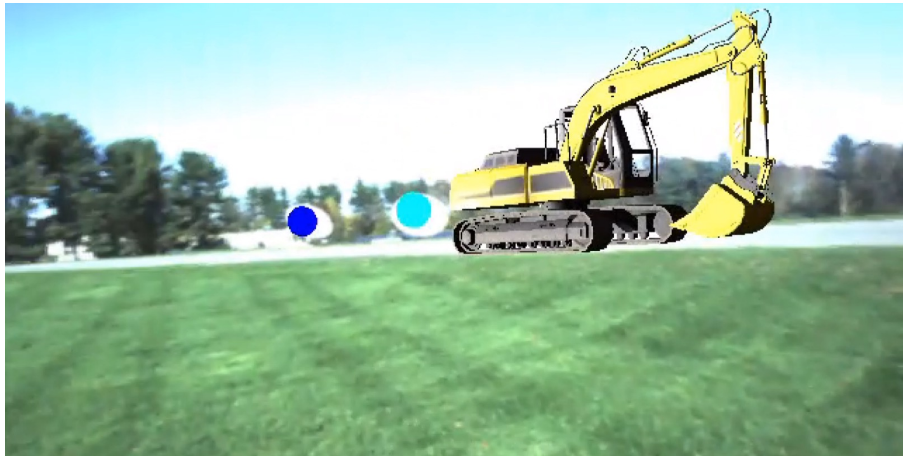
- Outdoor Augmented Reality (AR) **insertions with little to no drift**

+ Geo-located icons need to appear in correct location on AR display

+ Our goal: **Precise Global Location & Orientation Estimation** for compelling AR

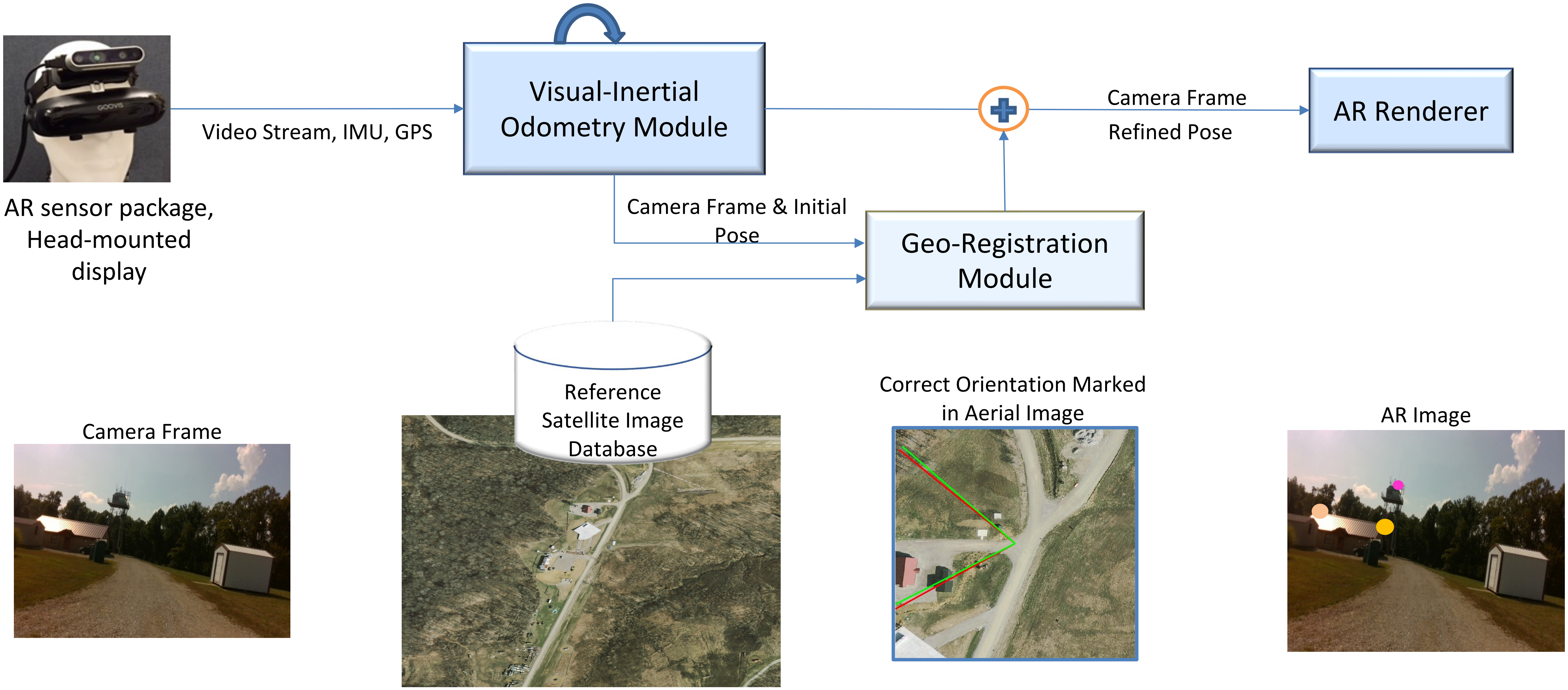
+ GPS and Magnetometer challenged situation





M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Geo-Alignment for estimating pose for Augmented Reality



M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Handling Real Sequences for AR application

- Benchmark data sets (CVUSA, CVACT) used for training neural network have 360 deg. Panorama for ground imagery
- Real world sequences often may be collected with cameras with smaller field of views (e.g. Real Sense has a 70 degree field of view)

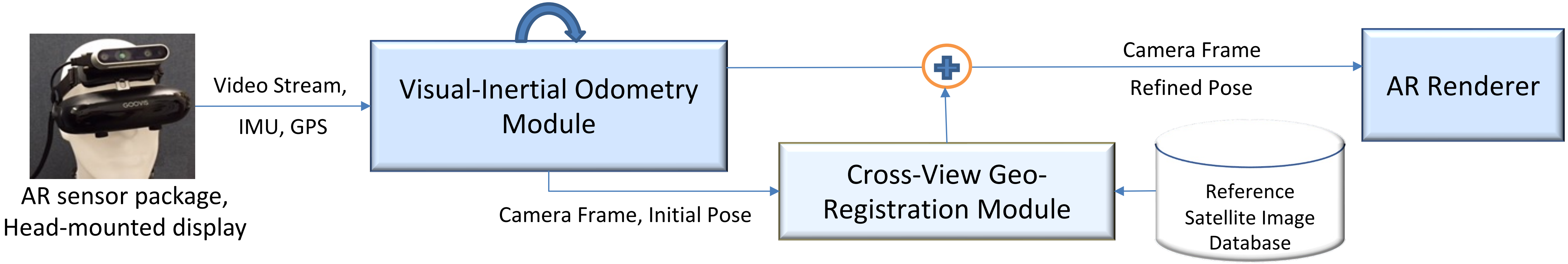
- To handle real world images, we do the following steps:
  - Fine tune the neural network with orientation loss and smaller field of view training data
  - Explore both transformer and CNN models.
    - CNN's have advantage to transformers that you can use different size images
    - With CNN, you can train with panorama data sets and fine tune on smaller field of view data
    - CNNs are also more efficient to run on edge processors
  - For pose update while moving: Develop methods for combining information from moving block of frames to get an effective wider field of view data for ground to air matching.
  - For cold start situation: Build panoramas, where user can just rotate in place to collect imagery for panorama. Use constructed panorama for air-ground matching
  - Confidence metrics on when to use the estimated solution

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Continuous sequence of frames and geo-registration

In the previous slides, we discussed approach for geo-registration from a single image.

- Works reasonably well with panorama images. However, when the camera FoV is small, a single frame have limited context and the estimate based on a single frame is not reliable/stable for AR/ navigation application.



<u>Approach for continuous Sequences of video frames </u>- Single Query-based Approach is extended to **using continuous frames with estimated Pose** from Visual Odometry to provide a **high-confidence and stable estimate.**

- Similarity scores for each orientation/position accumulated over a sequence using relative pose between frames.
- The approach can be used for both providing a cold-start and continuous refinement.
- Outlier Rejection: (i) Ratio Test based on the 1st and 2nd matching scores (ii) Field of View Coverage

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Experiments on Real-World Navigation Sequences

Ortho-Image

Ground-Truth

Predicted

Polar Transformed Ortho-Image

Qualitative Results on a sequence of 100 frames from Sequence from Johnson County, IN

Video Frames

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Experiments on Real-World Navigation Sequences

Collected Navigation Sequences

- Multiple sets of navigation sequences collected in different places across U.S., i.e., Mercer County, NJ; Prince William County, VA; Johnson County, IN.
- To create ground-truth., differential GPS and magnetometer are used as additional sensors.

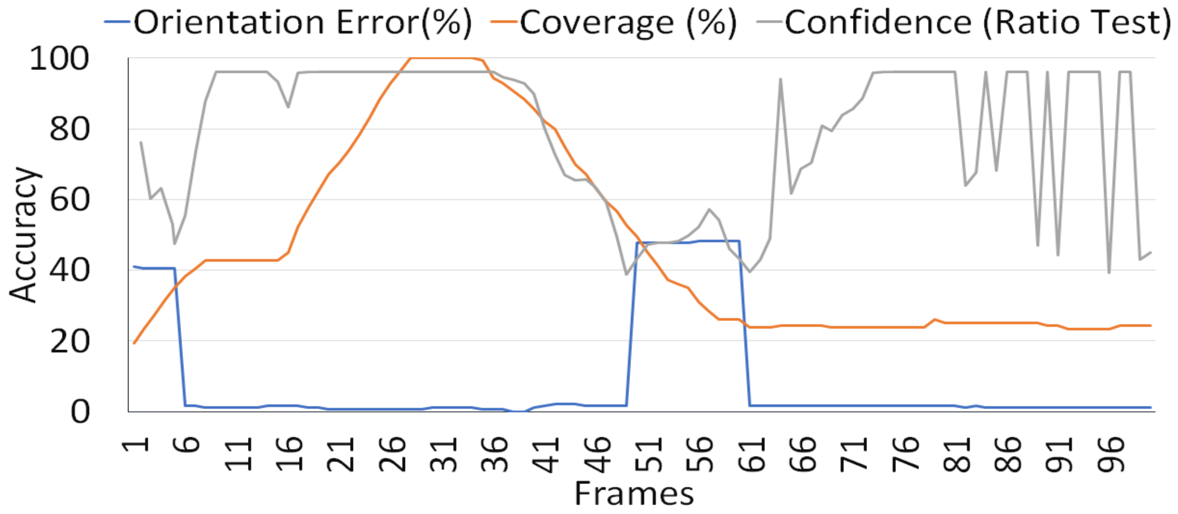


Fig. Orientation Estimation on a 100 frames from Set2, Sequence from Johnson County, IN. Last 20 frames are used in estimation.

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

Orientation Estimation (Accuracy within 2 Deg.) Results

FoV Coverage	Any	120	180
Set 1, Mercer County, New Jersey			
Ours, trained on CVACT /CVUSA	0.60	0.64	0.69
Ours (finetuned on nav seq.)	0.83	0.88	0.94
Set 2, Johnson County, Indiana			
Ours trained on CVACT /CVUSA	0.61	0.61	0.71
Ours (finetuned on nav seq.)	0.68	0.73	0.85

* Accuracy reported w/o considering outlier rejection based on Ratio-Test.

- As Field-of-View (FoV) Coverage increases, Error decreases.
- Confidence based on ratio test is very effective in avoiding most false positives.

Experiments on Real-World Navigation Sequences

Systems	RMS Error	Median Error	90th Percentile
GPS and Magnetometer available for the whole sequence.			
Nav. System	2.15	1.59	3.10
GPS and Magnetometer available for the whole sequence. Cross-View Geo-Registration Module is also used.			
Nav. System + Cross-View Geo-Reg. Module	2.08	1.48	3.08
GPS Challenged Case (Only an initial position estimate available). Magnetometer not available.			
Nav. System + Cross-View Geo-Reg. Module	2.51	1.89	3.79

Video

- Comparable results even in GPS and Magnetometer denied case.

M. Niluthpol et.al. Cross-View Visual Geo-Localization for Outdoor Augmented Reality, IEEE VR 2023

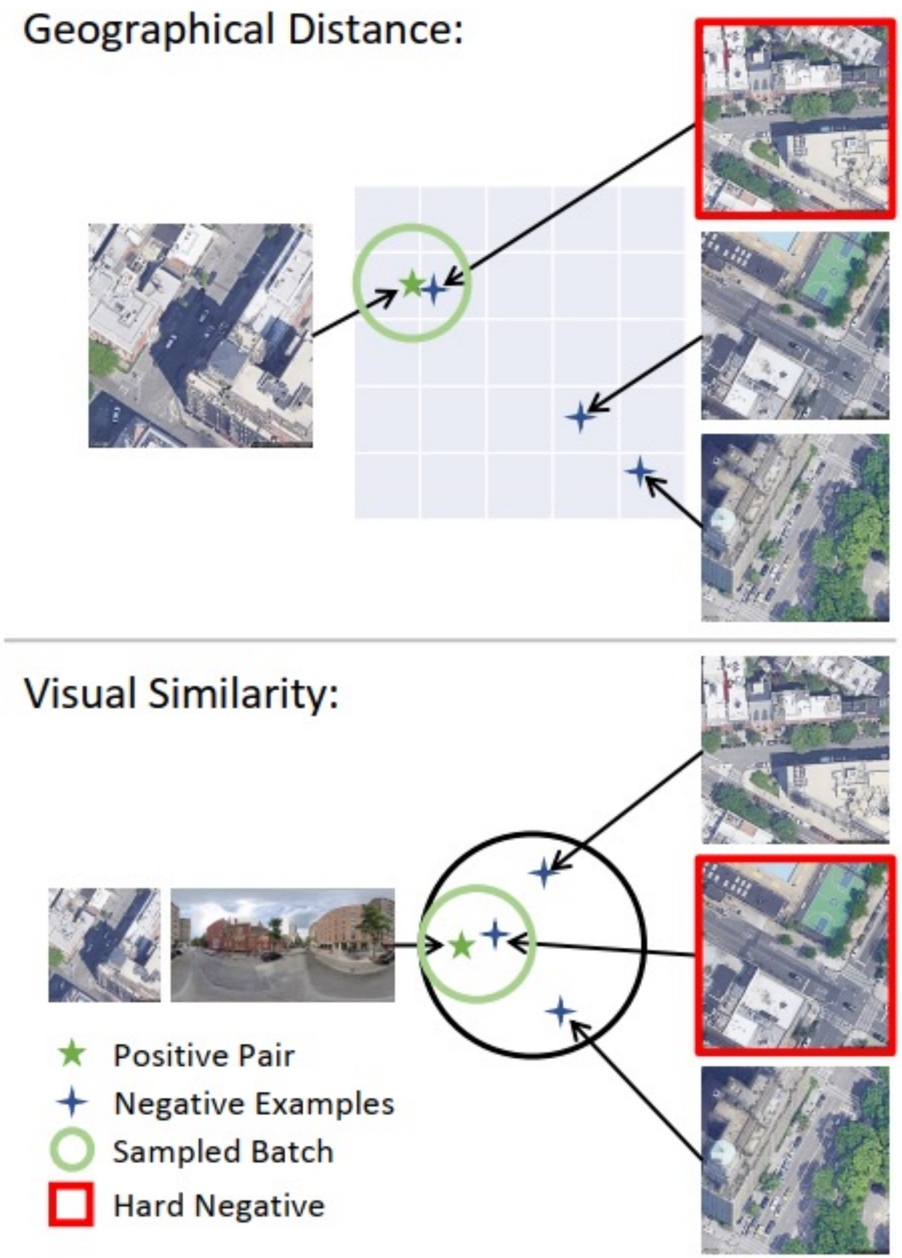
Sample4Geo: Hard Negative Sampling for Cross-Localization

3 key innovations

- ConvNext Model: Instead of transformers
  - A ConvNet for 2020's, https://arxiv.org/abs/2201.03545
  - Advantage: Can be used with images of different sizes
- Sampling hard negatives
  - Near Neighbor Sampling
  - Dynamic Sampling based on visual similiarity
- Symmetric InfoCE Loss: exploits all available negatives in the batch, as opposed to triplet loss

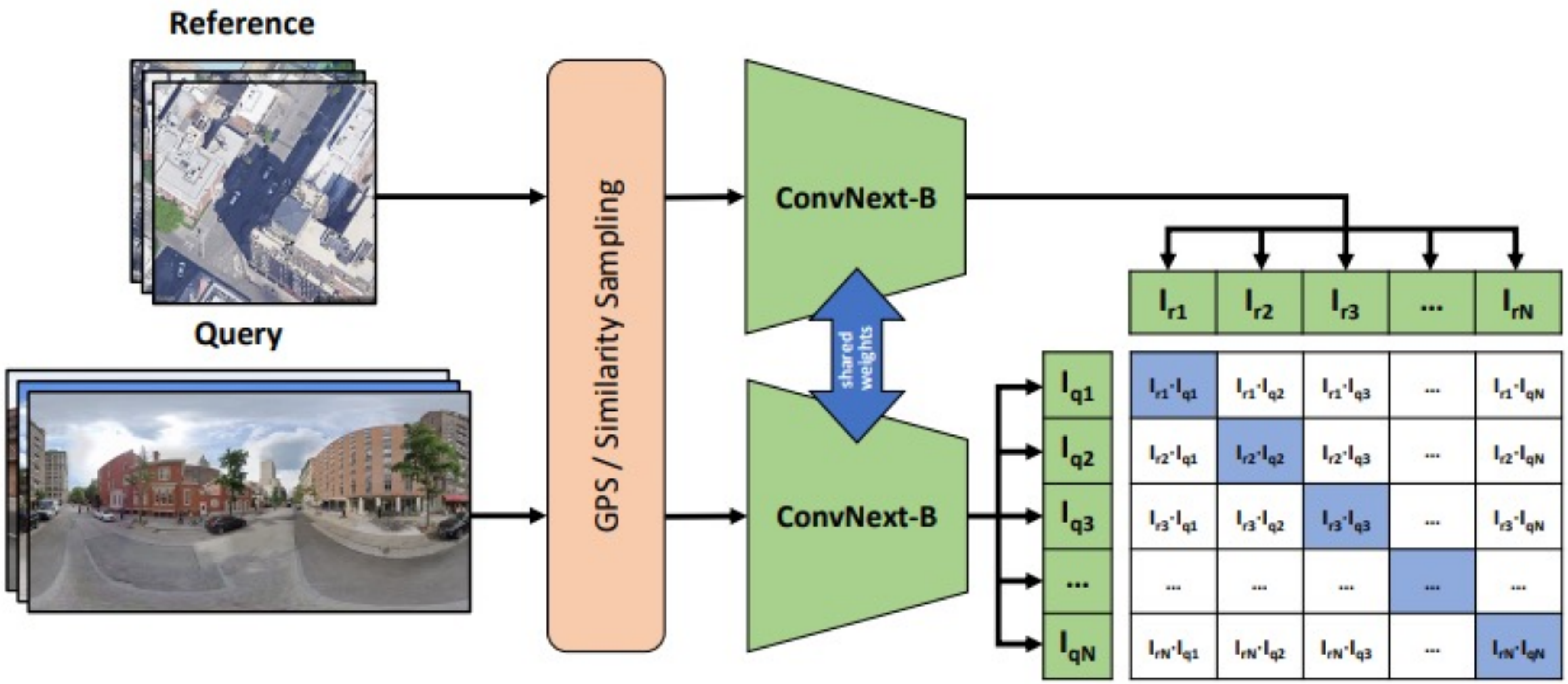
$$\mathcal{L}(q, R)_{\text{InfoNCE}} = -\log \frac{\exp(q \cdot r_{+}/\tau))}{\sum_{i=0}^{R} \exp(q \cdot r_i/\tau))}$$

q denotes an encoded street-view, the so-called query, and R is a set of encoded satellite images called references. Only one positive ri , namely r+ matches to q.



F. Deuser., et.al (2023): Sample4Geo: Hard Negative Sampling for Cross-view Geo-localization, https://arxiv.org/pdf/2303.11851.pdf

Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localization



- Uses an off-the-shelf ConvNeXt-B and mine hard negatives based on our GPS and visual similarity sampling. The InfoNCE loss is used in a symmetric fashion to learn discriminative features in both view directions.
- The same parameters are used for both ground image and overhead reference image

F. Deuser., et.al (2023): Sample4Geo: Hard Negative Sampling for Cross-view Geo-localization, https://arxiv.org/pdf/2303.11851.pdf

Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localization

Approach	CVUSA				CVACT Val				CVACT Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
LPN [19]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
SAFA† [20]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
TransGeo [24]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR [40]	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
GeoDTR†	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	**98.74**
SAIG-D [25]	96.08	98.72	99.22	99.86	89.21	96.07	97.04	98.74	-	-	-	-
SAIG-D† [25]	96.34	99.10	99.50	99.86	89.06	96.11	97.08	**98.89**	67.49	89.39	92.30	96.80
Ours	**98.68**	**99.68**	**99.78**	**99.87**	**90.81**	**96.74**	**97.48**	98.77	**71.51**	**92.42**	**94.45**	98.70

Table 1: **Quantitative comparison between our approach and state-of-the-art approaches on CVUSA [34] and CVACT [10].** † denotes which models are using the polar transformation for their satellite input as a pre-processing technique.

F. Deuser., et.al (2023): Sample4Geo: Hard Negative Sampling for Cross-view Geo-localization, https://arxiv.org/pdf/2303.11851.pdf

Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localization

Results on Vigor Dataset

Approach	R@1	R@5	R@10	R@1%	Hit Rate
SAME					
SAFA† [20]	33.93	58.42	68.12	98.24	36.87
TransGeo [24]	61.48	87.54	91.88	99.56	73.09
SAIG-D [25]	65.23	88.08	-	**99.68**	74.11
Ours	**77.86**	**95.66**	**97.21**	99.61	**89.82**
CROSS					
SAFA† [20]	8.20	19.59	26.36	77.61	8.85
TransGeo [24]	18.99	38.24	46.91	88.94	21.21
SAIG-D [25]	33.05	55.94	-	94.64	36.71
Ours	**61.70**	**83.50**	**88.00**	**98.17**	**69.87**

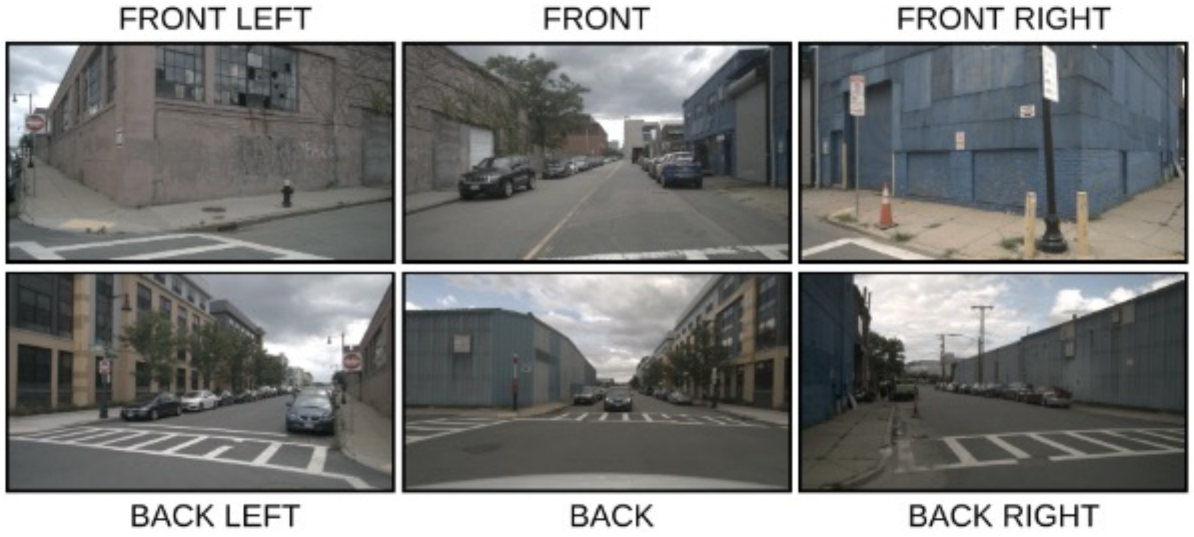
Ours is Sample4Geo method.
Quantitative comparison between Sample4Geo approach and other methods on VIGOR dataset

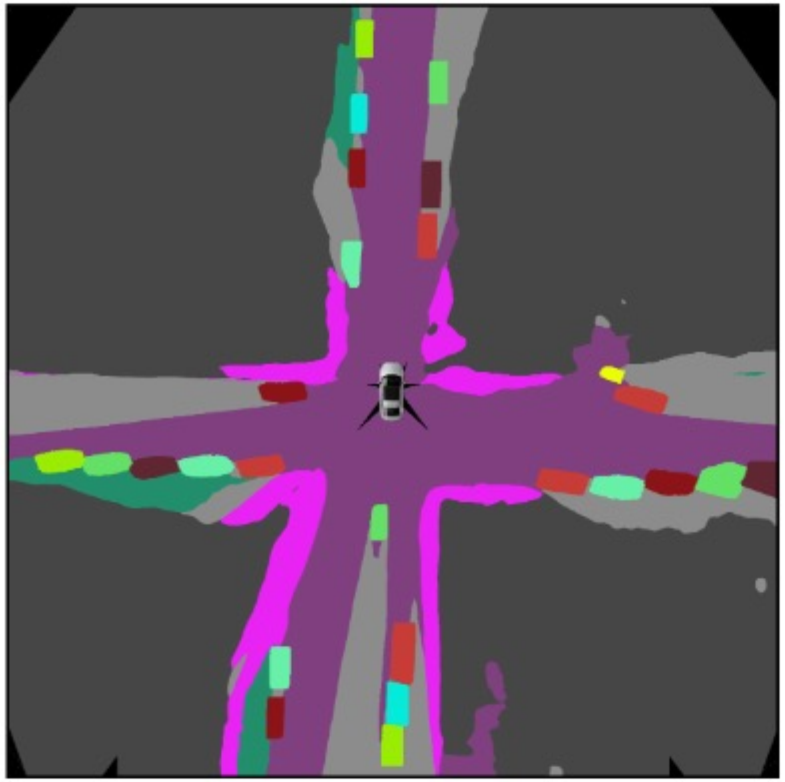
F. Deuser., et.al: Sample4Geo: Hard Negative Sampling for Cross-view Geo-localization, https://arxiv.org/pdf/2303.11851.pdf

Agenda

- Introduction to Problem
- Cross view matching approaches: Match 2D ground images to 2D overhead reference
  - Invariant features/ Project reference image to ground view
  - Project ground image to overhead view/ Bird's eye view
- Cross view matching: 2D ground images to 2.5 D overhead reference (2D reference image with 3D point cloud or terrain data)
- Cross modal matching: 2D ground images to LIDAR reference

Birds Eye View Maps and Images (BEV)





Bird's-Eye-View (BEV) maps portray the scene around an platform as if it were viewed by a bird overflying the platform.

Nikhil Gosala, Abhinav Valada, "Bird's-Eye-View Panoptic Segmentation Using Monocular Frontal View Images", *IEEE Robotics and Automation Letters (RA-L)*, 2022.

A taxonomy of algorithms for perspective view to bird's eye view

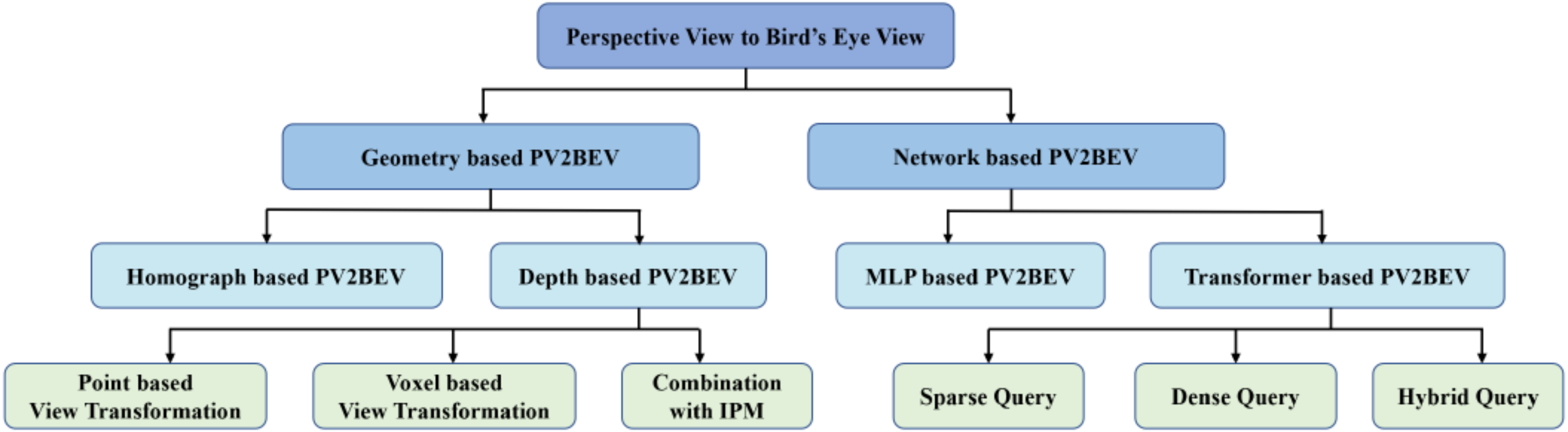
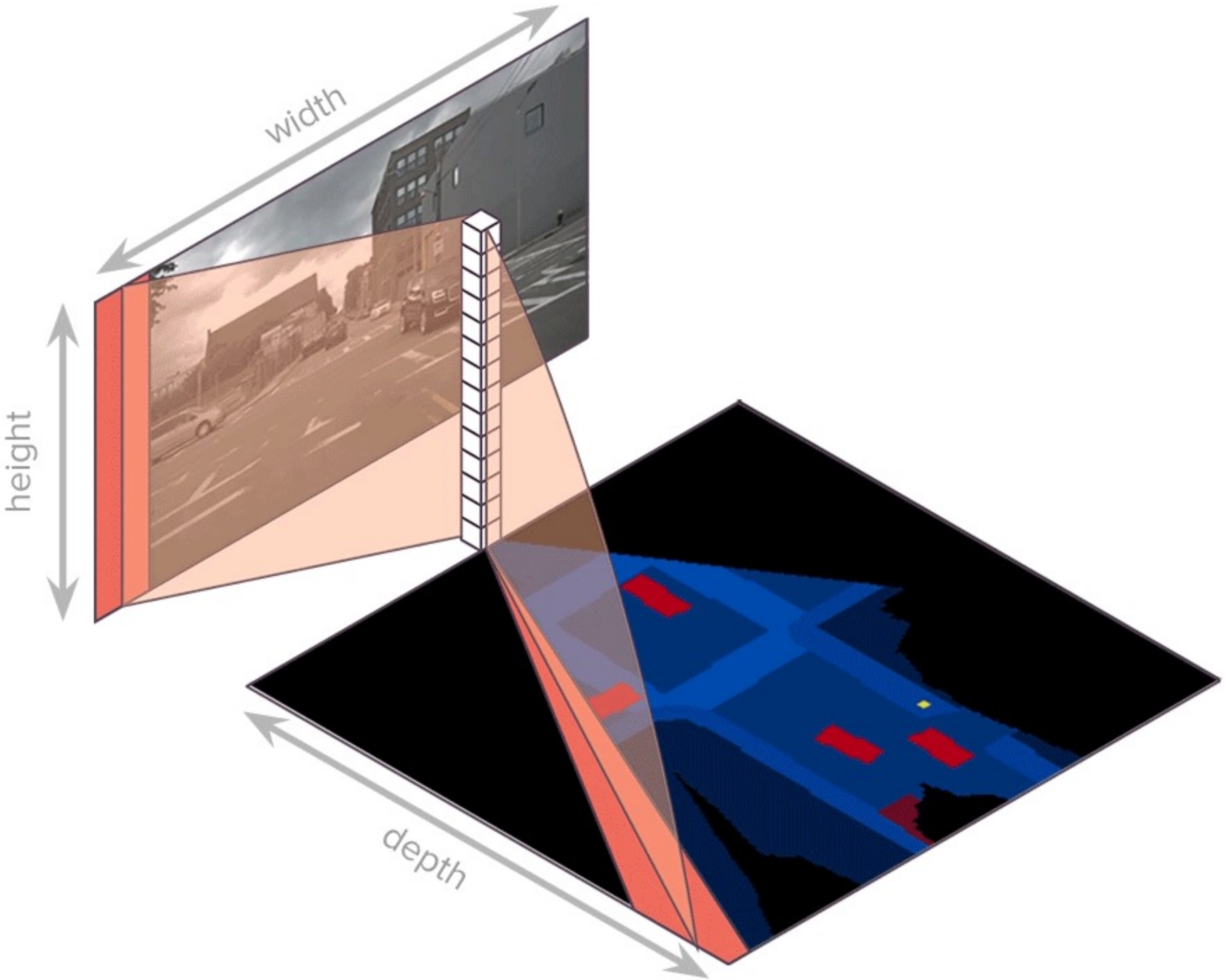


Figure from Ma. et.al., "Vision-Centric BEV Perception: A Survey", https://arxiv.org/abs/2208.02797

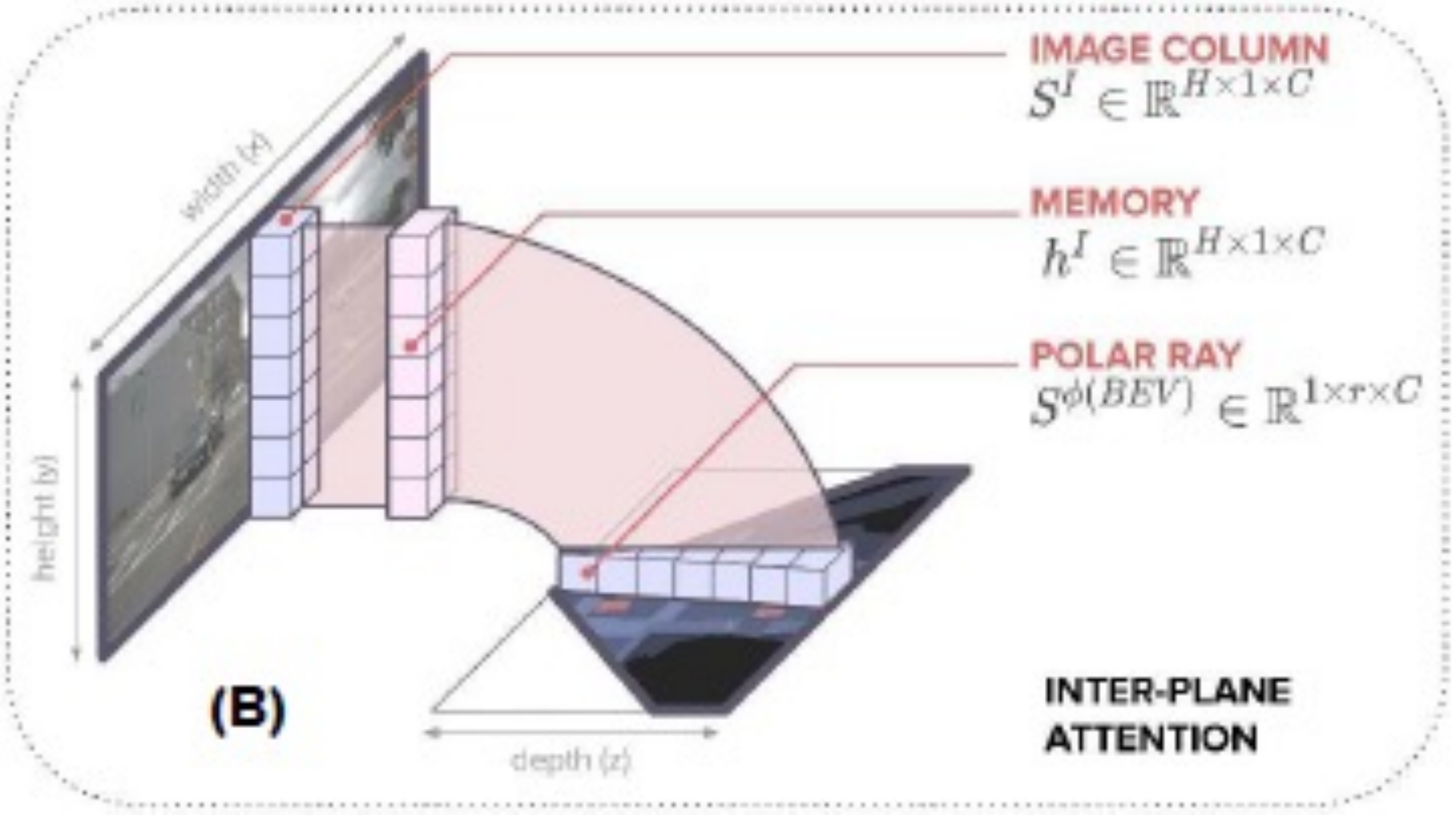
Translating images into maps (ICRA 2022)



- Approach the image-to-world transformation as a translation problem
- Assume a 1-1 correspondence between vertical scan lines in the image and BEV polar rays
- Transformer is convolutional in the horizontal direction, making efficient use of data when training
  - Our transformer encoder and decoder use the same set of projection matrices for every sequence-to-sequence translation, giving it a structure that is convolutional along the x-axis
- Achieved state of the Art results for instantaneous mapping on nuScenes

A.Saha et.al, Translating images into maps, ICRA 2022, https://arxiv.org/abs/2110.00966

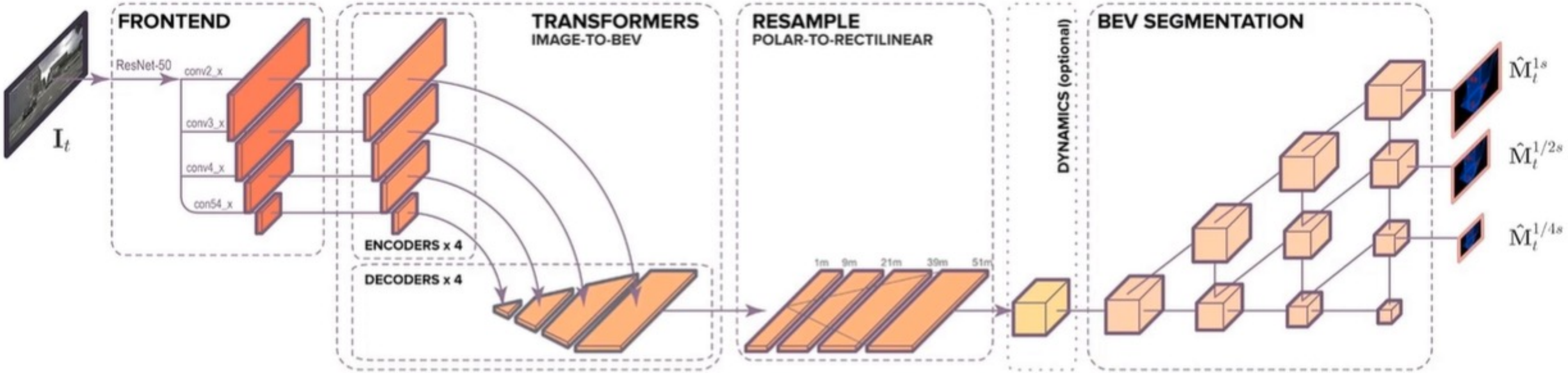
Inter-plane attention mechanism



The inter-plane attention mechanism. In the attention-based model, vertical scan lines in the image are passed one by one to a transformer encoder to create a 'memory' representation which is decoded into a BEV polar ray.

A.Saha et.al, Translating images into maps, ICRA 2022, https://arxiv.org/abs/2110.00966

Architecture



The Frontend extracts spatial features at multiple scales. Encoder-decoder transformers translate spatial features from the image to BEV. An optional Dynamics Module uses past spatial BEV features to learn a spatiotemporal BEV representation. A BEV Segmentation Network processes the BEV representation to produce multi-scale occupancy grids.

A.Saha et.al, Translating images into maps, ICRA 2022, https://arxiv.org/abs/2110.00966

Translating images into maps (ICRA 2022)

Method	Drivable	Crossing	Walkway	Carpark	Bus	Bike	Car	Cons.Veh.	Motorbike	Trailer	Truck	Ped.	Cone	Barrier	Mean
VED [6]	54.7	12.0	20.7	13.5	0.0	0.0	8.8	0.0	0.0	7.4	0.2	0.0	0	4.0	8.7
VPN [2]	58.0	27.3	29.4	12.3	20.0	4.4	25.5	4.9	5.6	**16.6**	17.3	7.1	4.6	10.8	17.5
PON [8]	60.4	28.0	31.0	18.4	20.8	9.4	24.7	12.3	7.0	**16.6**	16.3	8.2	5.7	8.1	19.1
STA-S [10]	71.1	31.5	32.0	28.0	22.8	14.6	34.6	10.0	7.1	11.4	18.1	7.4	5.8	10.8	21.8
Our Spatial	**72.6**	**36.3**	**32.4**	**30.5**	**32.5**	**15.1**	**37.4**	**13.8**	**8.1**	15.5	**24.5**	**8.7**	**7.4**	**15.1**	**25.0**
STA-ST [10]	70.7	31.1	32.4	**33.5**	29.2	12.1	36.0	12.1	**8.0**	13.6	22.8	8.6	6.9	14.2	23.7
Our Spatiotemp.	**74.5**	**36.6**	**35.9**	31.3	**32.8**	**14.7**	**39.7**	**14.2**	7.6	**13.9**	**26.3**	**9.5**	**7.6**	**14.7**	**25.7**

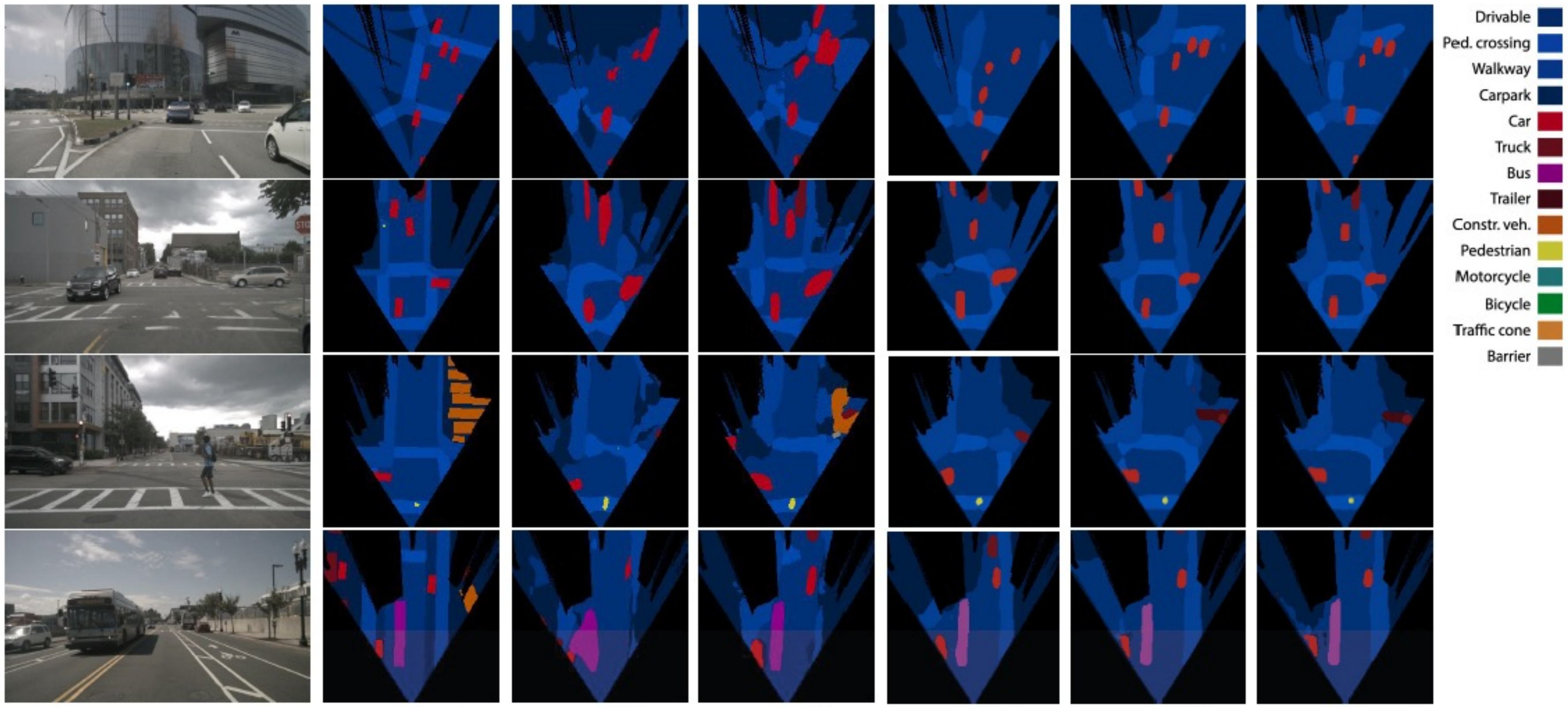
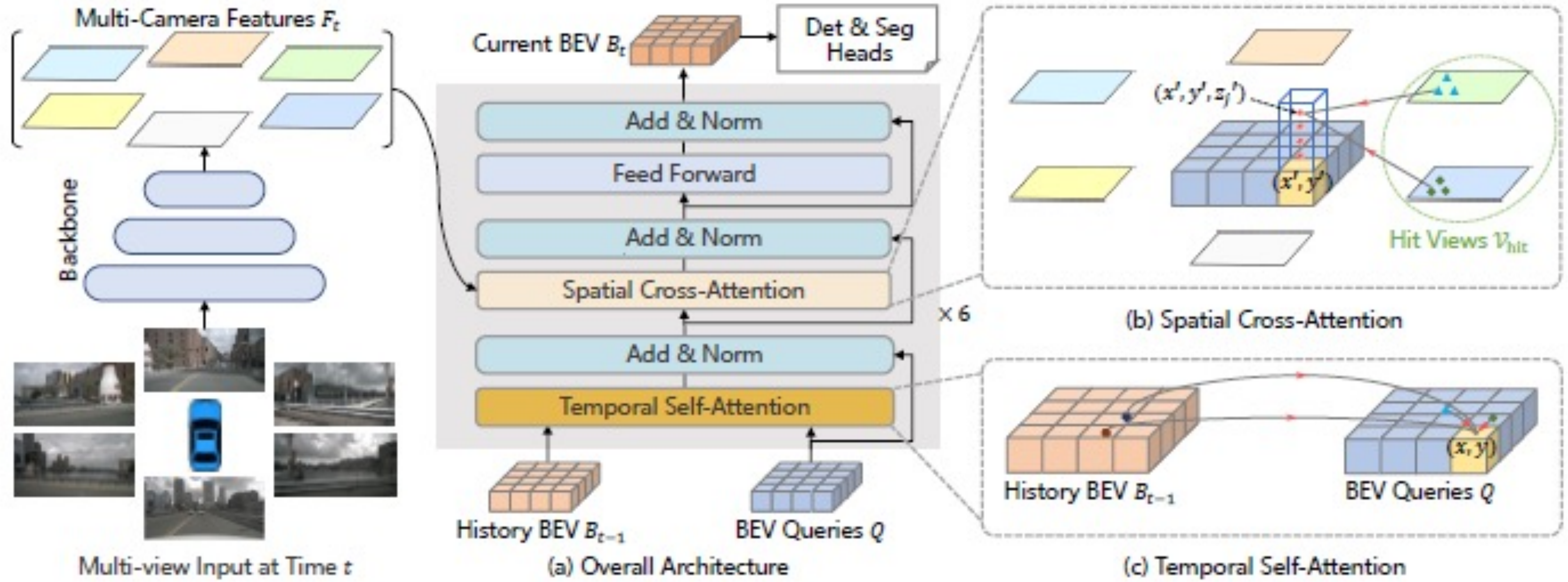


Image | Ground truth | VPN [2] | PON [8] | STA-S [10] | Our spatial | Our spatiotemporal

Qualitative results of A. Saha et.al. ICRA,2022) on the nuScenes validation set.

A.Saha et.al, Translating images into maps, ICRA 2022, https://arxiv.org/abs/2110.00966

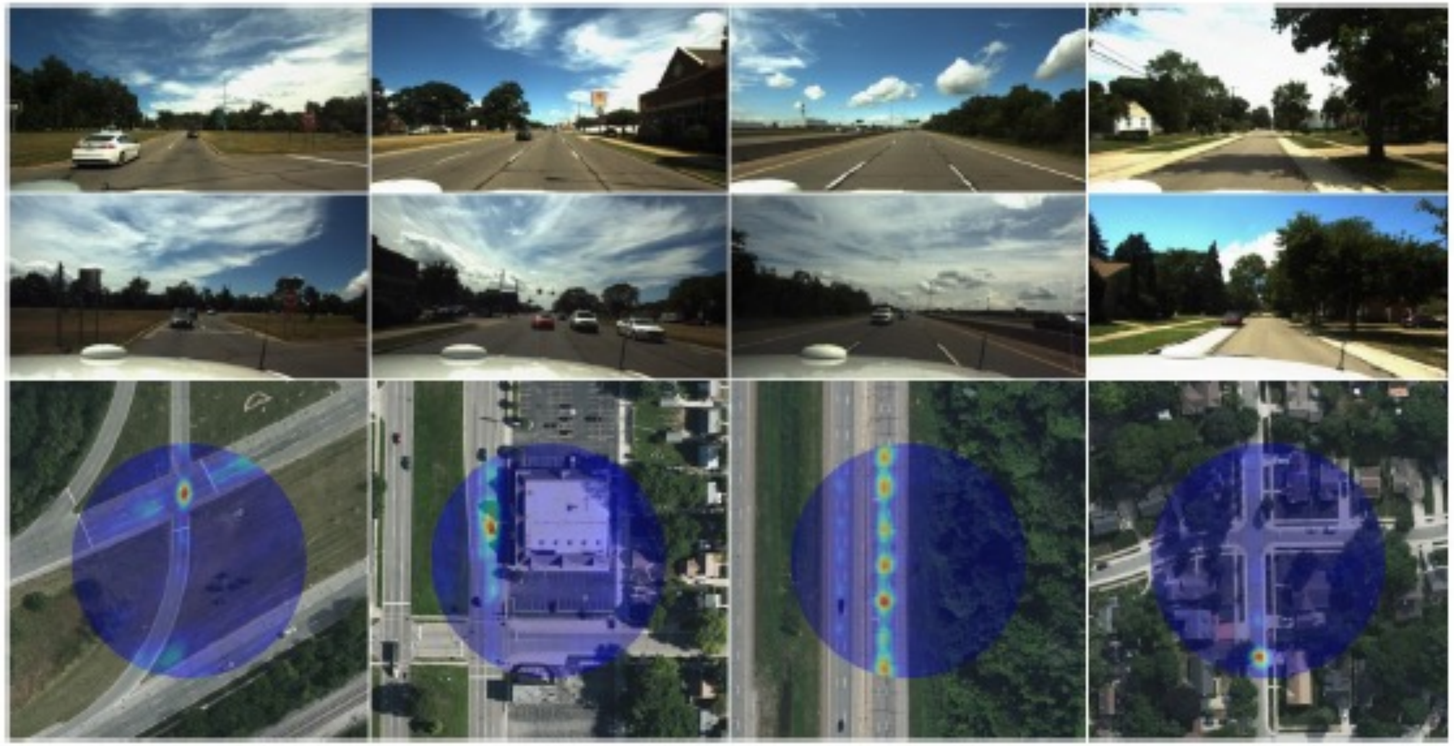
BEVFORMER



Overall architecture of BEVFormer. (a) The encoder layer of BEVFormer contains grid-shaped BEV queries, temporal self-attention, and spatial cross-attention. (b) In spatial cross-attention, each BEV query only interacts with image features in the regions of interest. (c) In temporal self-attention, each BEV query interacts with two features: the BEV queries at the current timestamp and the BEV features at the previous timestamp.

Z. Li., et.al., BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers, ECCV 2022, https://arxiv.org/abs/2203.17270

Uncertainty-aware Vision-based Metric Cross-view Geo-localization



Probability distributions for the vehicle position predicted by our model which matches the vehicle's surround camera images with an aerial image.

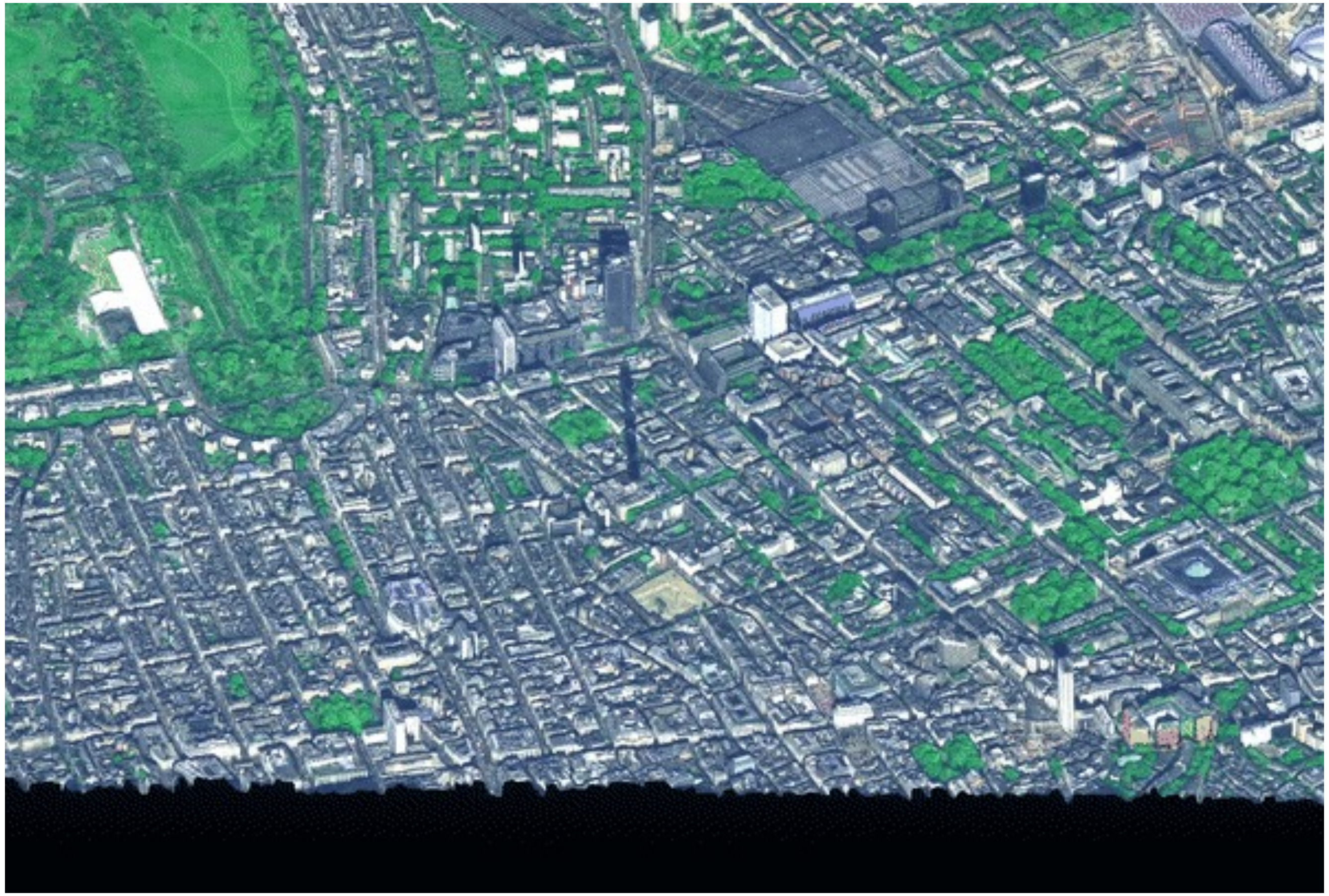
- The first and second rows show the front and back cameras in the Ford AV dataset.
- The last row shows the aerial image with the search region in the center and driving direction pointing upwards.
- Blue and red color refer to low and high probability predicted by our model.

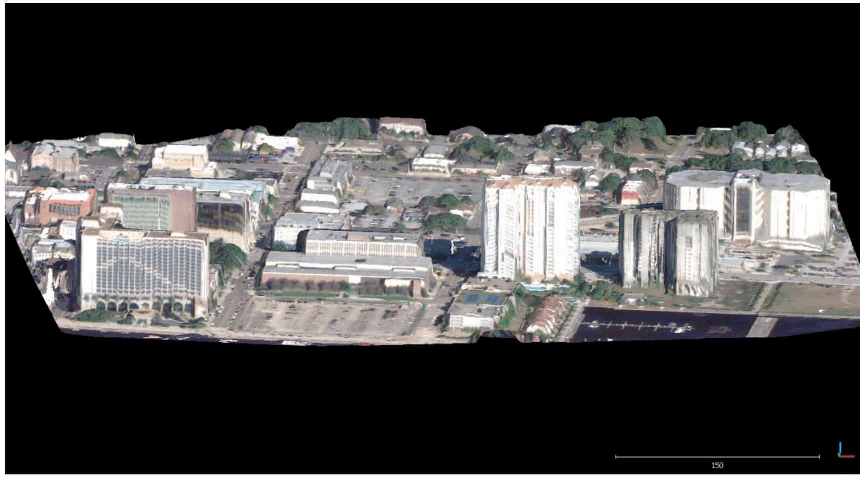
Fervers et.al., Uncertainty-aware Vision-based Metric Cross-view Geo-localization, CVPR 2023, https://arxiv.org/abs/2211.12145

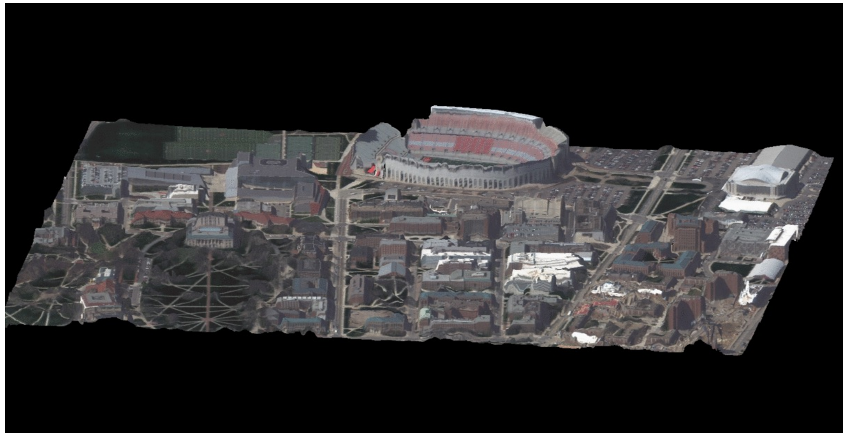
Agenda

- Introduction to Problem
- Cross view matching approaches: Match 2D ground images to 2D overhead reference
  - Invariant features/ Project reference image to ground view
  - Project ground image to overhead view/ Bird's eye view
- Cross view matching: 2D ground images to 2.5 D overhead reference (2D reference image with 3D point cloud or terrain data)
- Cross modal matching: 2D ground images to LIDAR reference

Multi-view 3D Reconstruction from Satellite images

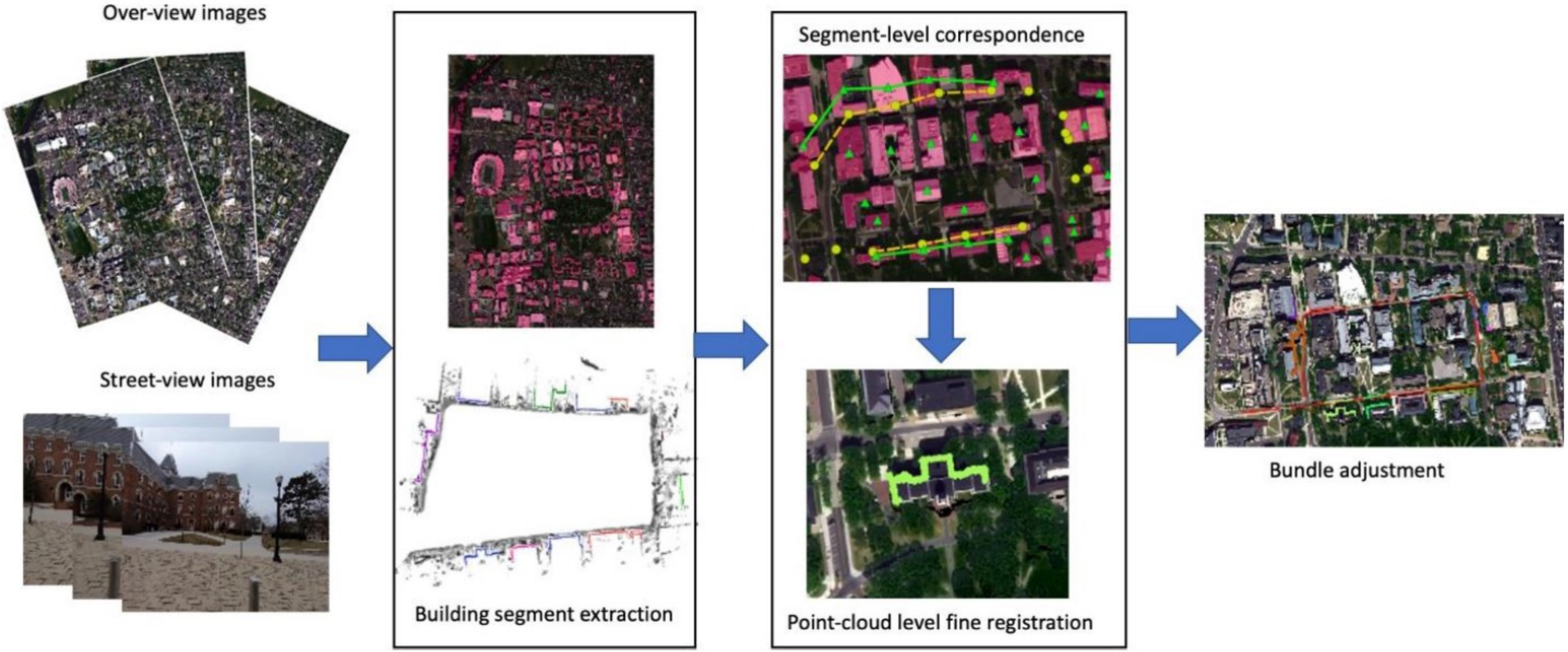






Work from Prof. Rongjun Qin's Lab., Dept. of Civil and Geodetic Eng., OSU
A Critical Analysis of Satellite Stereo Pairs for Digital Surface Model Generation and A Matching Quality Prediction Model – ISPRS J.2019.
A Unified Framework of Bundle Adjustment and Feature Matching for High-Resolution Satellite Images – PE&RS, 2021
Disparity refinement in depth discontinuity using robustly matched straight lines for digital surface model generation – IEEE JSTARS, 2019
Automated 3D recovery from very high resolution multi-view images Overview of 3D recovery from multi-view satellite image – ASPRS conf.

Cross-view (Air-Ground) Registration



Work from Prof. Rongjun Qin's Lab., Dept. of Civil and Geodetic Eng., OSU

A Graph-Matching Approach for Cross-view Registration of Over-view 1 and Street-view based Point Clouds. **ISPRS J. 2021**

Geo-alignment

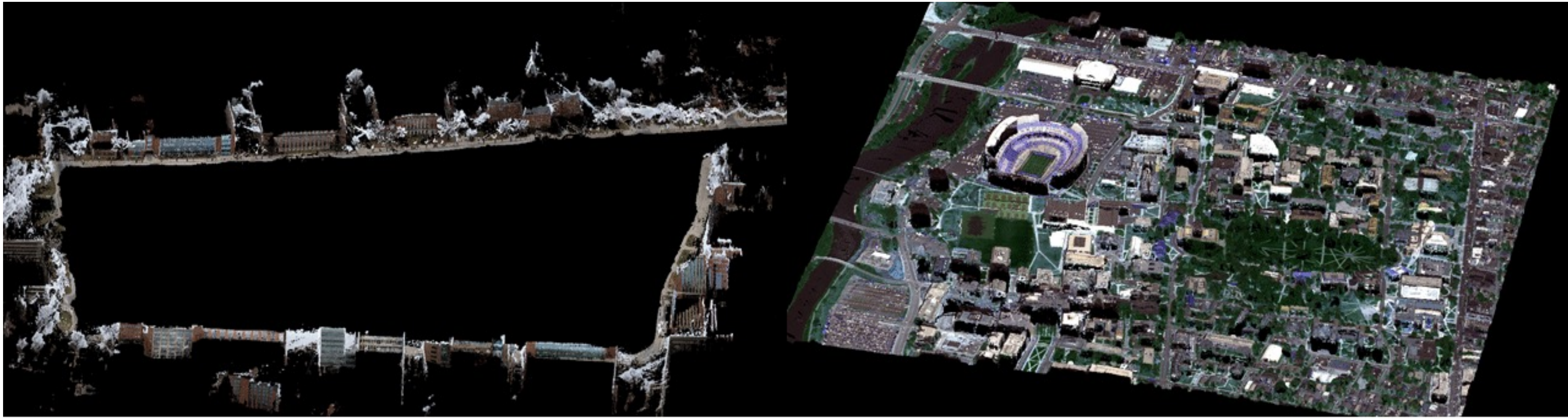
Go-pro videos
GPS tags available

Satellite data

2010-9-25 2011-10-8 2013-8-6
2014-6-6 2015-4-17

Work from Prof. Rongjun Qin's Lab., Dept. of Civil and Geodetic Eng., OSU
A Graph-Matching Approach for Cross-view Registration of Over-view 1 and Street-view based Point Clouds. **ISPRS J. 2021**

Cross-view (Air-Ground) Registration

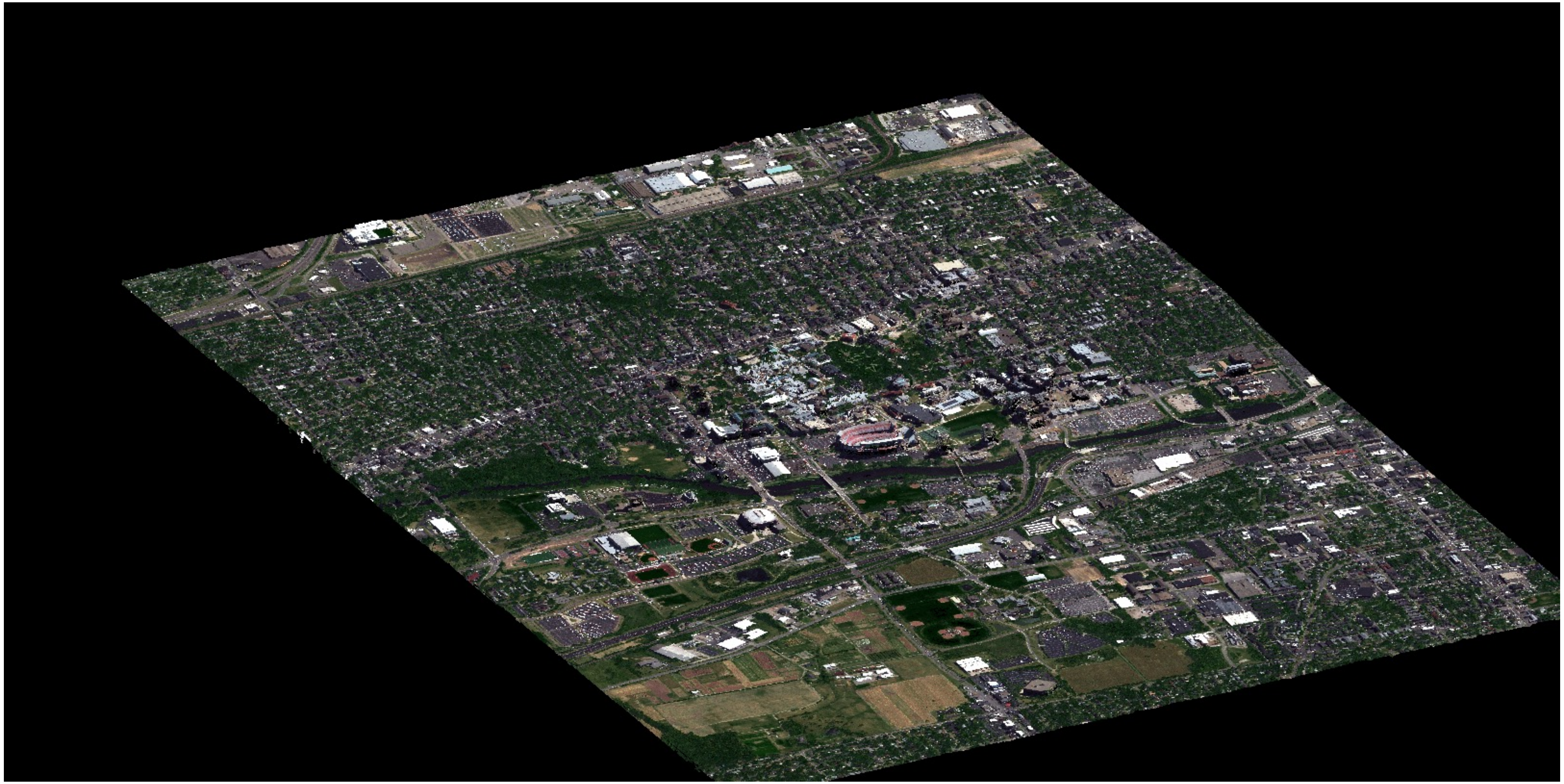


Streetview Pointcloud   Overview Pointcloud

Work from Prof. Rongjun Qin's Lab., Dept. of Civil and Geodetic Eng., OSU

A Graph-Matching Approach for Cross-view Registration of Over-view 1 and Street-view based Point Clouds. **ISPRS J. 2021**

Cross-view (Air-Ground) Registration Results (integrating Satellite and Ground point clouds)



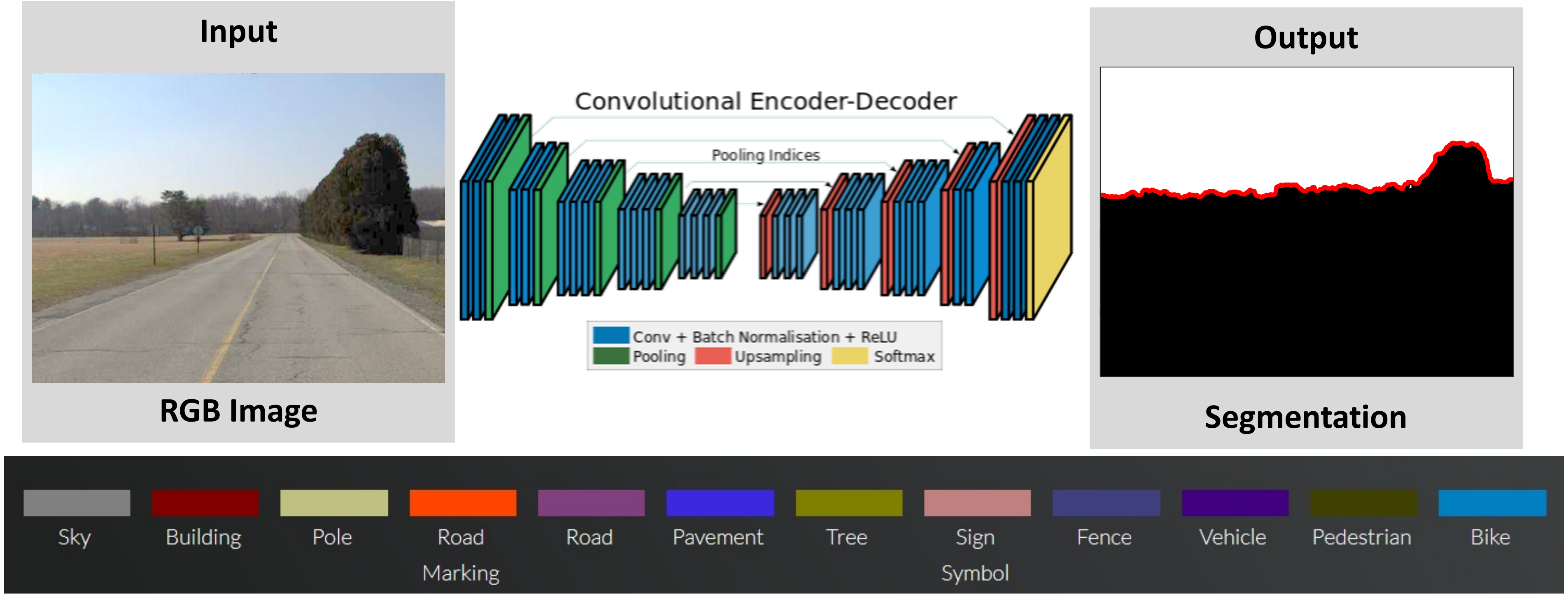
Work from Prof. Rongjun Qin's Lab., Dept. of Civil and Geodetic Eng., OSU
A Graph-Matching Approach for Cross-view Registration of Over-view 1 and Street-view based Point Clouds. **ISPRS J. 2021**

Agenda

- Introduction to Problem
- Cross view matching approaches: Match 2D ground images to 2D overhead reference
  - Invariant features/ Project reference image to ground view
  - Project ground image to overhead view/ Bird's eye view
- Cross view matching: 2D ground images to 2.5 D overhead reference (2D reference image with 3D point cloud or terrain data)
- Cross modal matching: 2D ground images to LIDAR reference

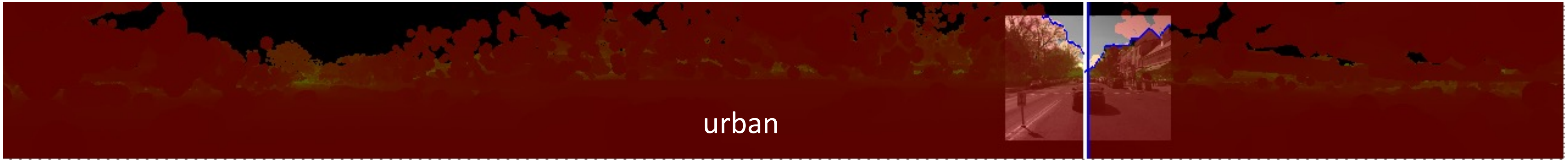
Geo-registration of ground imagery to overhead reference with 3D terrain

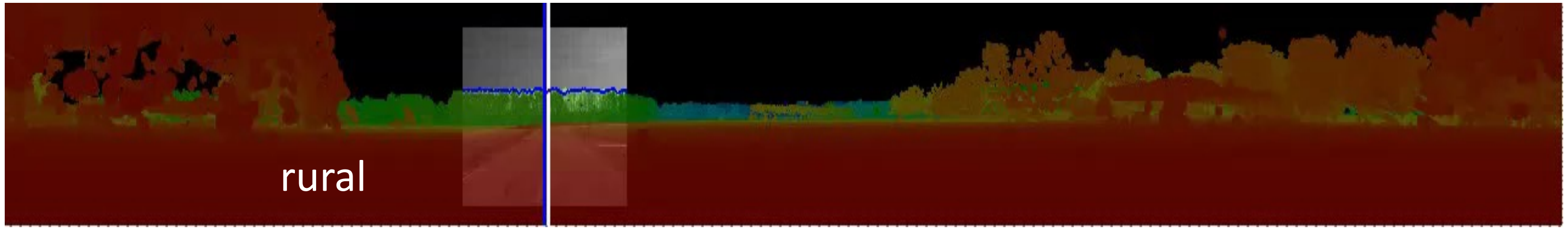
- Estimate absolute heading by matching skyline extracted from reference data to skyline visible in images
- Input image from dismount platform is processed using SegNet to extract skylines to generate an edge template.
- 3D terrain in reference data is processed to create skyline from reference.



Geo-Spatial Association – Semantic Geo-Registration

Perform 2D-3D geo-registration continuously between the input video frame and the matched LIDAR depth data. Below shows the computed global heading based on skyline matching (the estimated heading accuracy is 0.4970 degree.

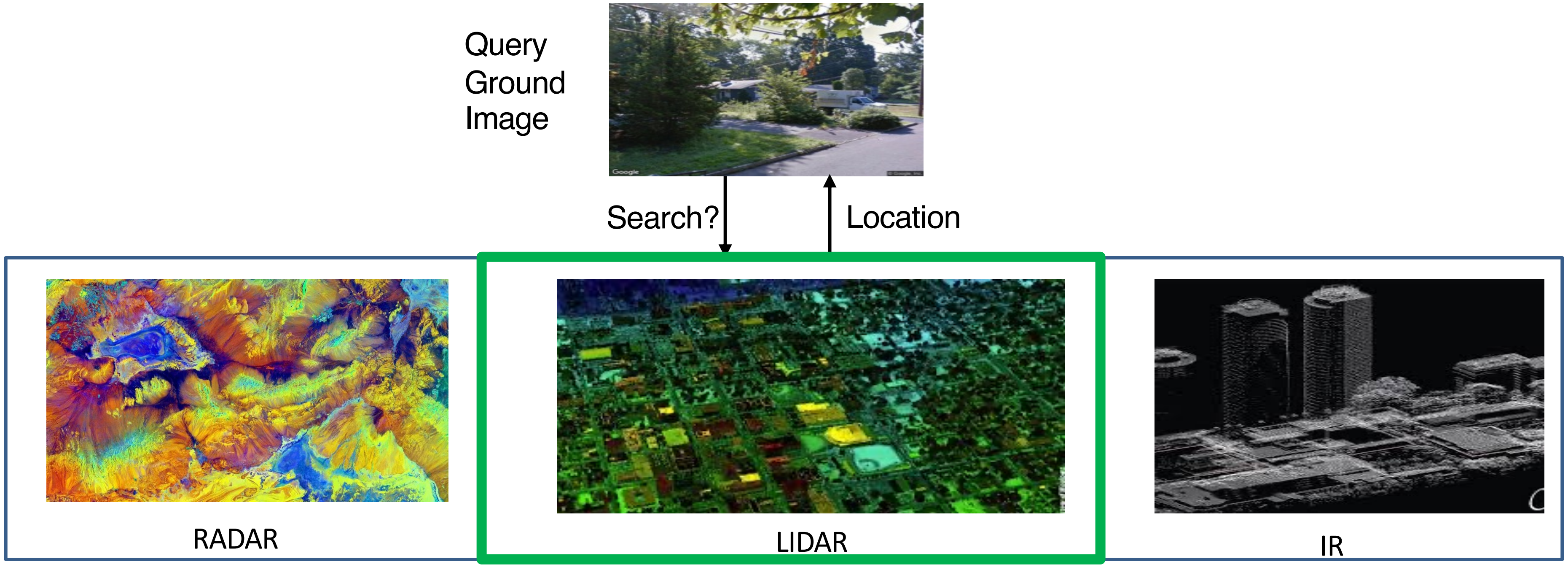




Han-Pang Chiu et al., **Augmented Reality Driving Using Semantic Geo-Registration**, *IEEE International Conference on Virtual Reality (VR)*, 2018.

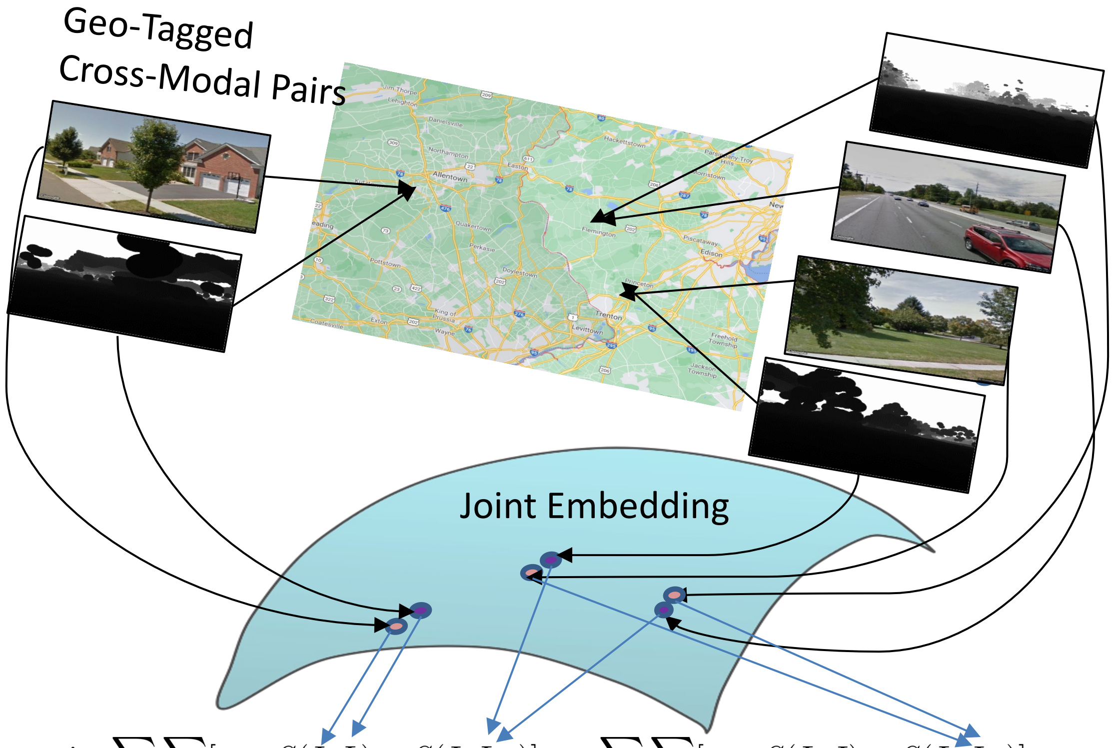
Cross-Modal Vision-based Geo-Localization

Cross-Modal Matching: Match to Geo-Tagged Cross-Modal Data



Cross-Modal Sources

Training Joint RGB-LIDAR Embedding



$$\min_{\theta} \sum_{I} \sum_{L^-} [\alpha - S(I,L) + S(I,L^-)]_+ + \sum_{L} \sum_{I^-} [\alpha - S(L,I) + S(L,I^-)]_+$$

(I, L) is a matching pair in the embedding. L^- is a non-matching lidar embedding for I and vice versa.

$[a]_+ = max(0, a)$

First Deep Learning based Method from Cross-Modal VL

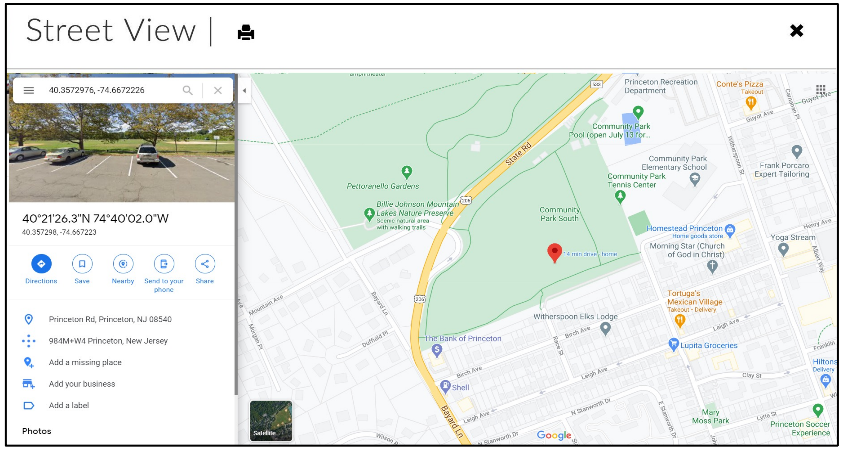
Joint RGB-LIDAR Embedding

- Cross-modal pairs closer in the geo-space should be closer in the embedding space

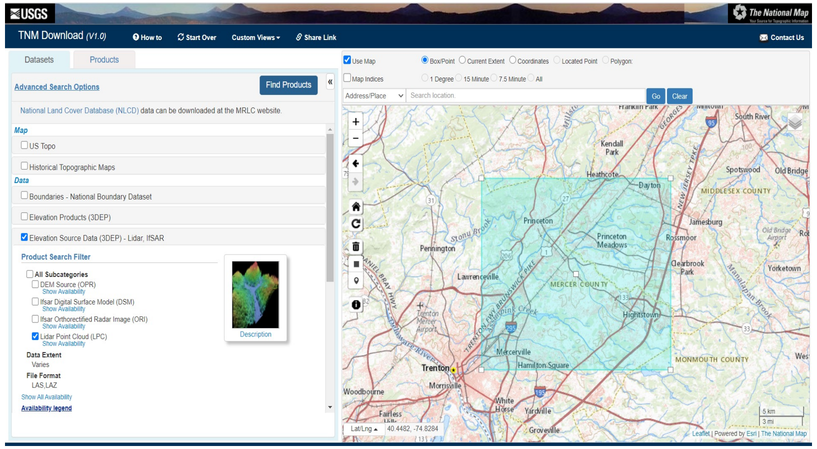
RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization
https://arxiv.org/abs/2009.05695

GRAL Dataset

Dataset Constructed with Automatically Collected Location-Coupled Cross-Modal Pairs



Images for locations using Google Street-View API



Render Images for locations from Lidar Point-Cloud (from USGS)

- About **550K cross-modal pairs** collected from $143\ \mathrm{km}^2$ area in NJ, USA.
- **Automatic collection** of pairs based on Location (Latitude, Longitude, Heading)

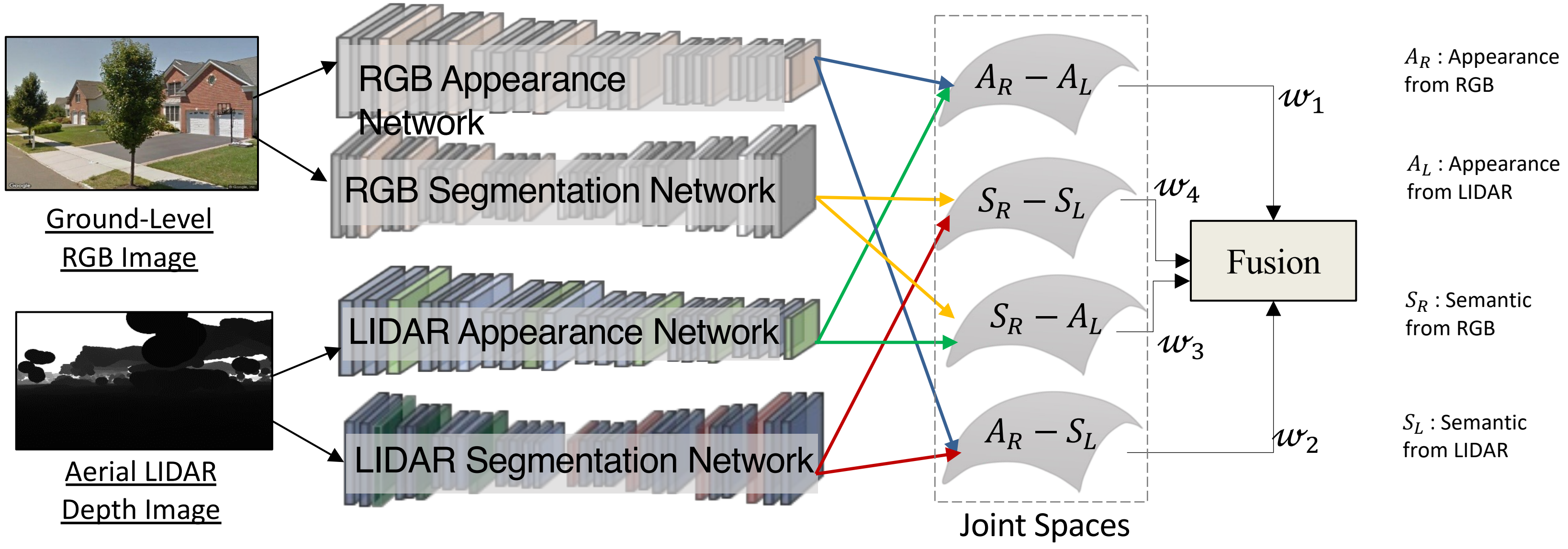
Dataset Available Online at https://github.com/niluthpol/RGB2LIDAR

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization
https://arxiv.org/abs/2009.05695

Fusion of Appearance and Semantic Cues

Both Appearance and Semantic Cues for Retrieval

- Matching across modalities exhibits large disparities in appearance.
- Higher-level scene information is generally better preserved across visual sensors.

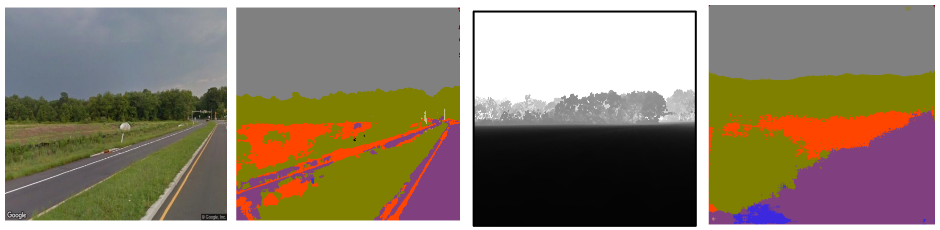


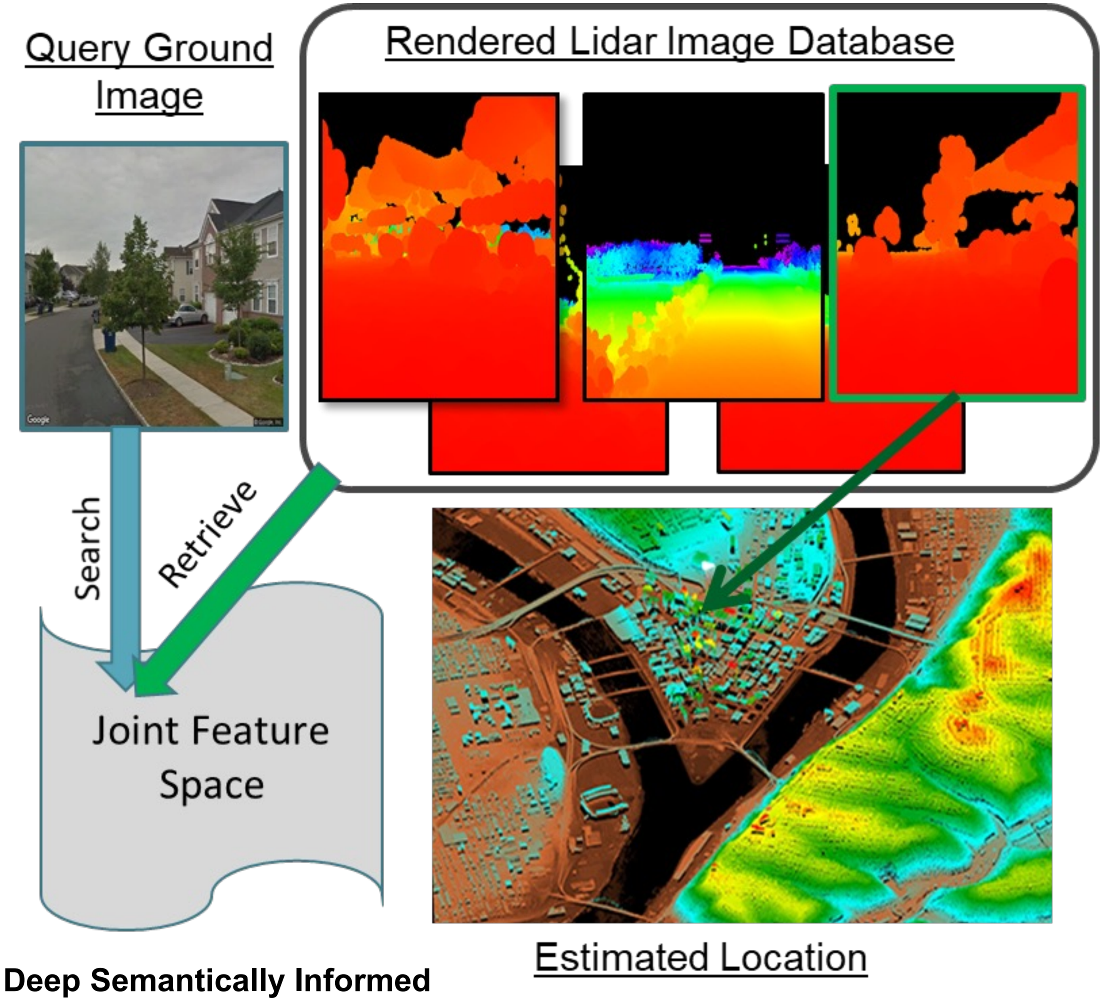
Mixture-of-Expert Model for Retrieval

Niluthpol C. Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar, **Deep Semantically Informed Cross-Modal Visual Localization**. ACM Multimedia (MM), 2020, Best Paper Finalist.

Cross-Modal Visual Localization

- Estimate position for a given **query ground image** by matching to a database of **aerial LIDAR geo-referenced data.**
  - We learn an embedding network to project visual features and lidar depth features into same joint space, that the embedded vectors of a image and lidar pair is closer for a correct match.
  - We supervise training and embedding of semantic labels on LIDAR depth data based on image semantic labels.
  - **The average median rank is 5 for our matching over 143 Km^2 database.**





Niluthpol C. Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar, **Deep Semantically Informed Cross-Modal Visual Localization**. ACM Multimedia (MM), 2020, Best Paper Finalist.

Summary

Presented Techniques for

- Cross view matching approaches: 2D ground images to 2D overhead reference
  - Invariant Features
  - Project reference image to ground view
  - Project ground image to overhead view/ Birds eye view
- Cross view matching: 2D ground images to 2.5D/ 3D reference
- Cross modal matching: 2D ground images to LIDAR reference

Questions

Cross-view matching talks in the afternoon

- **Cross-view and Cross-Modal Geo-localization:**
  a) 12.30 – 1.30 PM: Cross View and Cross-Modal Coarse Search and Fine alignment for Augmented Reality, Navigation and other applications, Rakesh (Teddy) Kumar, in-person.
  b) 1.30 PM – 2.30 PM: Toward Real-world Cross-view Geo-localization, Chen Chen/Sijie Zhu, in-person.
  c) 2.30 – 3.00 PM: Vision-based Metric Cross-view Geo-localization, Florian Fervers, in-person.

- **3:00 PM – 3:30 PM Coffee Break**

- **Cross-view and Cross-modal Geo-localization continuation:**
  a) 3.30 PM – 4.30 PM: Geometry-based Cross-view Geo-localization and Metric Localization for Vehicle, Yujiao Shi, in-person.
  b) 4.30 – 5.30 PM: Learning Disentangled Geometric Layout Correspondence for Cross-View Geo-localization Waqas Sultani, virtual.