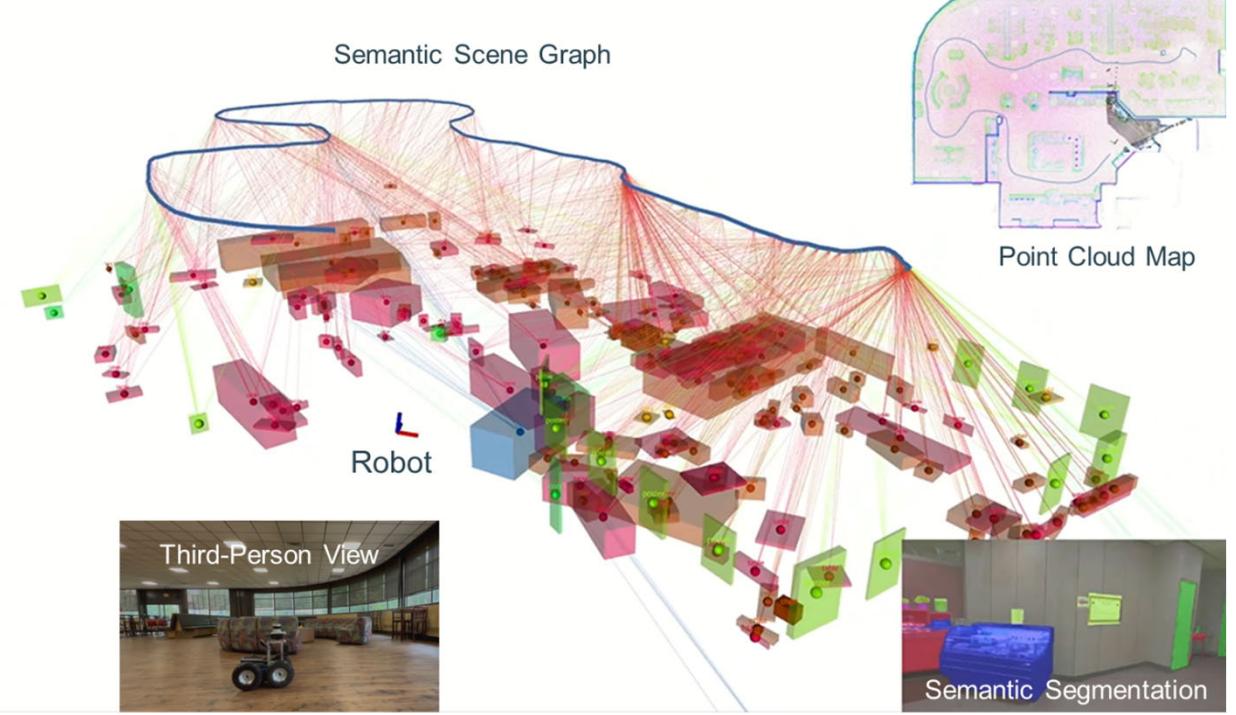


Semantically Guided Collaborative Navigation, 3D Mapping, Planning and Control for Unmanned Platforms

Rakesh (Teddy) Kumar
rakesh.kumar@sri.com

Center for Vision Technologies
SRI International



The robot continuously builds the point cloud map and segments objects from images (semantic segmentation) for incrementally generating the semantic scene graph.

Our Mission

SRI creates **WORLD-CHANGING SOLUTIONS**

making people safer, healthier,
and more productive

Legacy

Founded in 1946 by Stanford University, independent in 1970

Sarnoff Corp.(RCA Labs) founded in 1942, merged with SRI in 2011

Xerox Palo Alto Research Center (PARC), merged with SRI in 2023.



PARC

Research Focus



Human Augmentation

Intelligent interactive systems that augment human abilities



Automation & Infrastructure

Smart and secure systems that enhance capacity, capability and connectivity of businesses

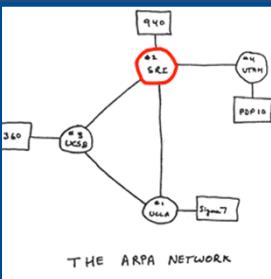


Healthcare

Technologies that improve patient outcomes and lower healthcare costs

SRI International®

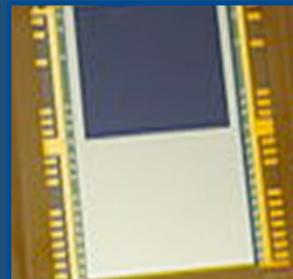
Legacy of Innovation



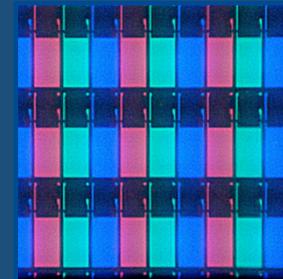
Original computer mouse



Liquid Crystal Display



1st CCD Devices



Thin Film Transistors



1st Virtual Private Network



1st Autonomous Robot



Pioneered Robotic surgery



Siri - 1st virtual assistant



1st Augmented Reality Broadcast



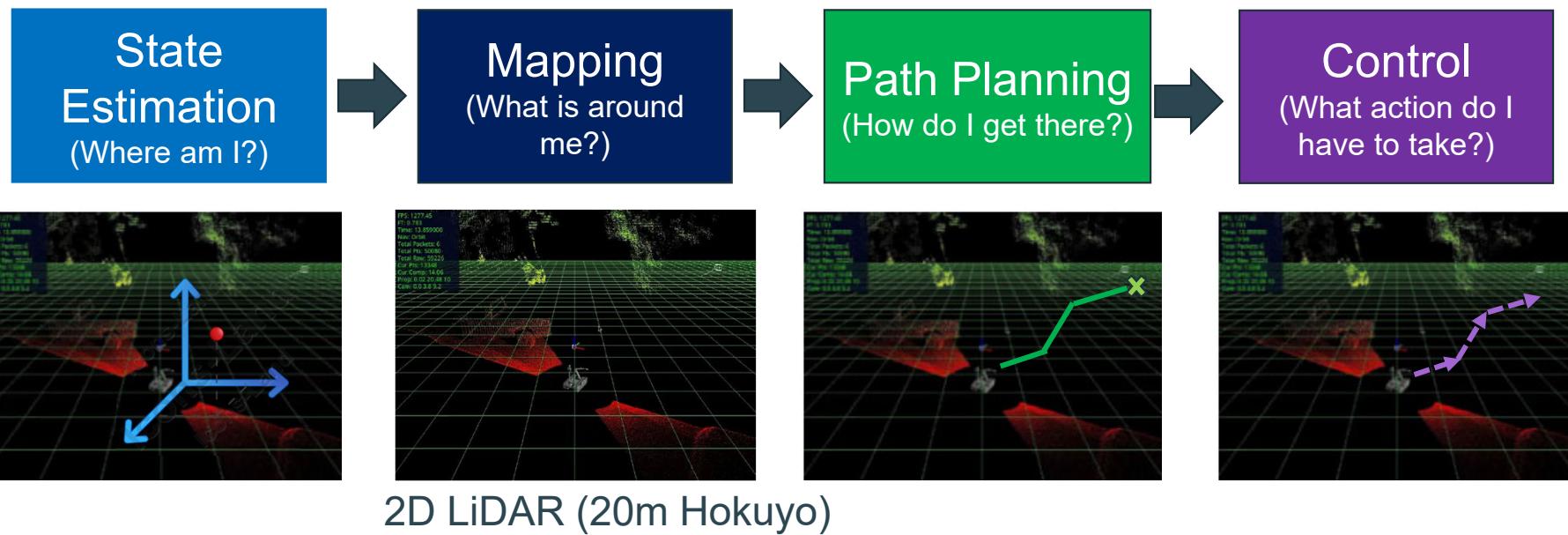
Established Network Intrusion Detection



Led HDTV Grand Alliance

Robotic Autonomy

- SRI CVT has a long history in robotic autonomy.
 - Robotic autonomy in large-scale environments is an important and challenging problem to many applications.



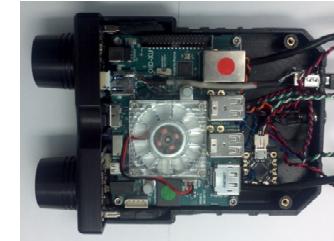
*<https://www.sri.com/hoi/shakey-the-robot/>

SRI Cam Slam System

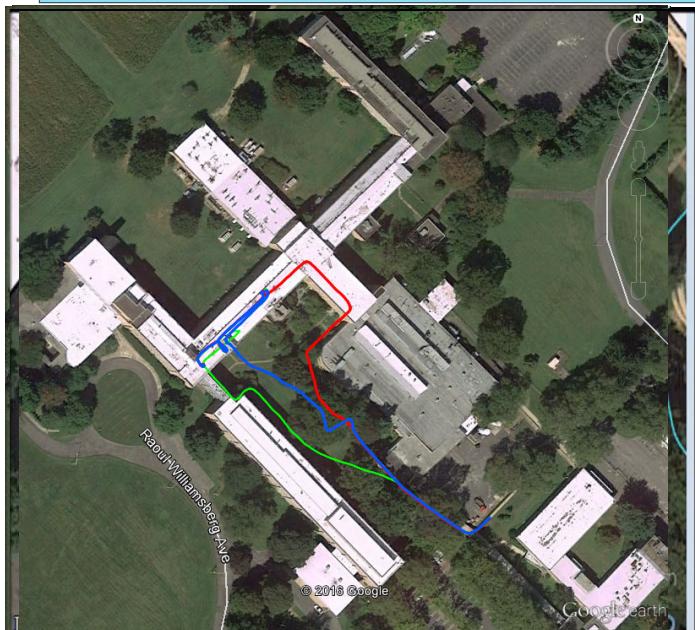
Navigation based on dynamic map creation and matching

Dynamically created maps allows the system to ensure low drift errors in subsequent runs

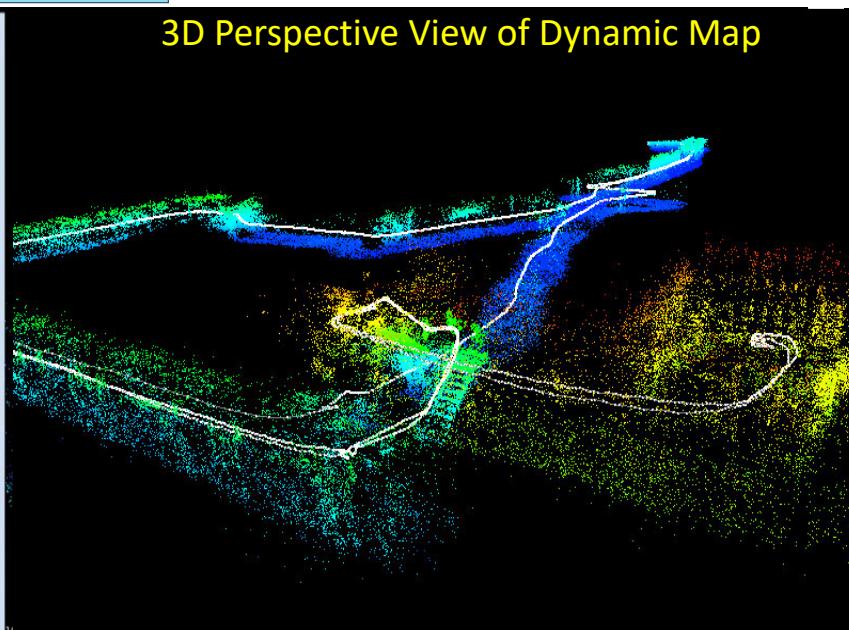
- On the fly Creation of Visual Maps
- Match to Dynamic Maps to reset Drift
- Create Maps Across Multiple Platforms or Runs



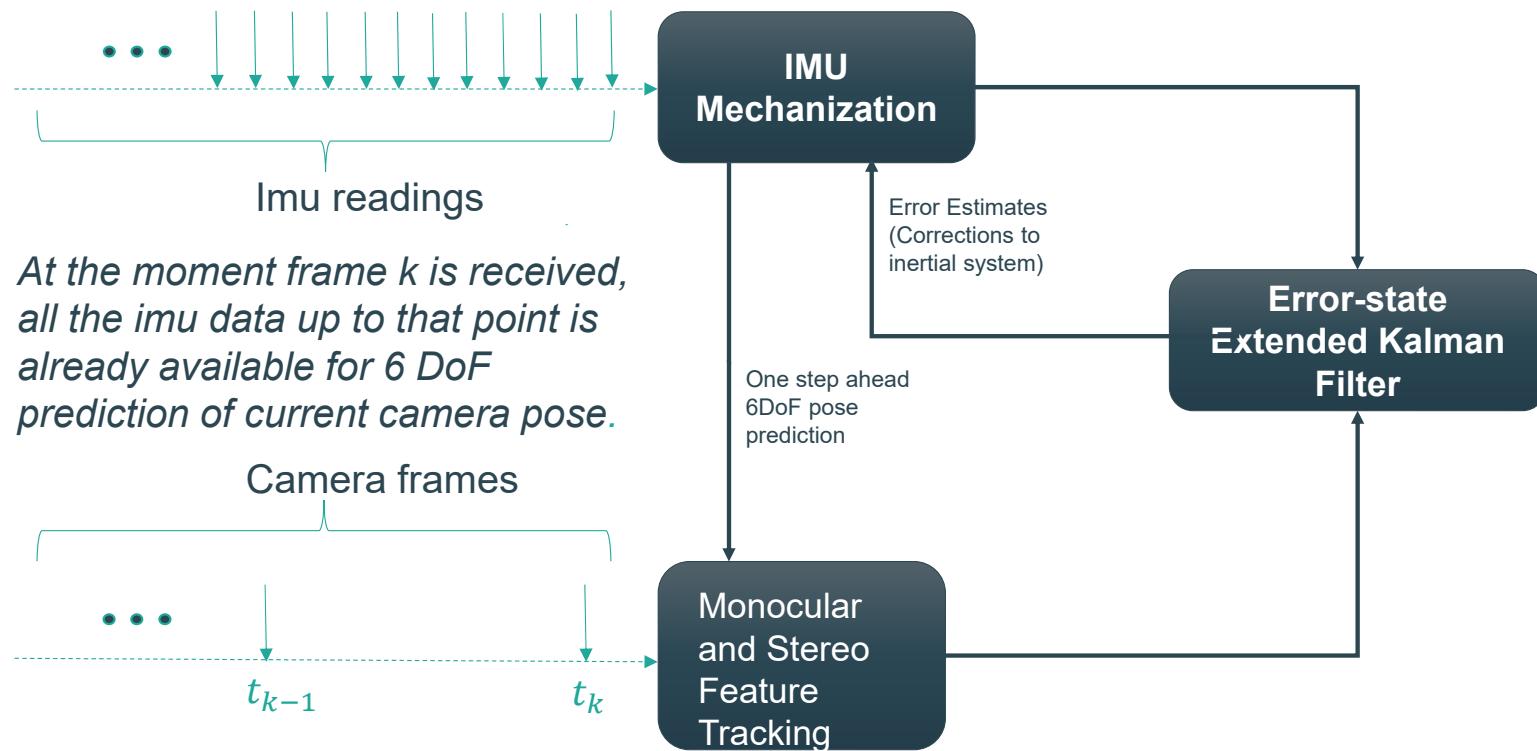
Navigation sensors: IMU, Cameras, GPS, Magnetometer, Barometer, Ranging Radios



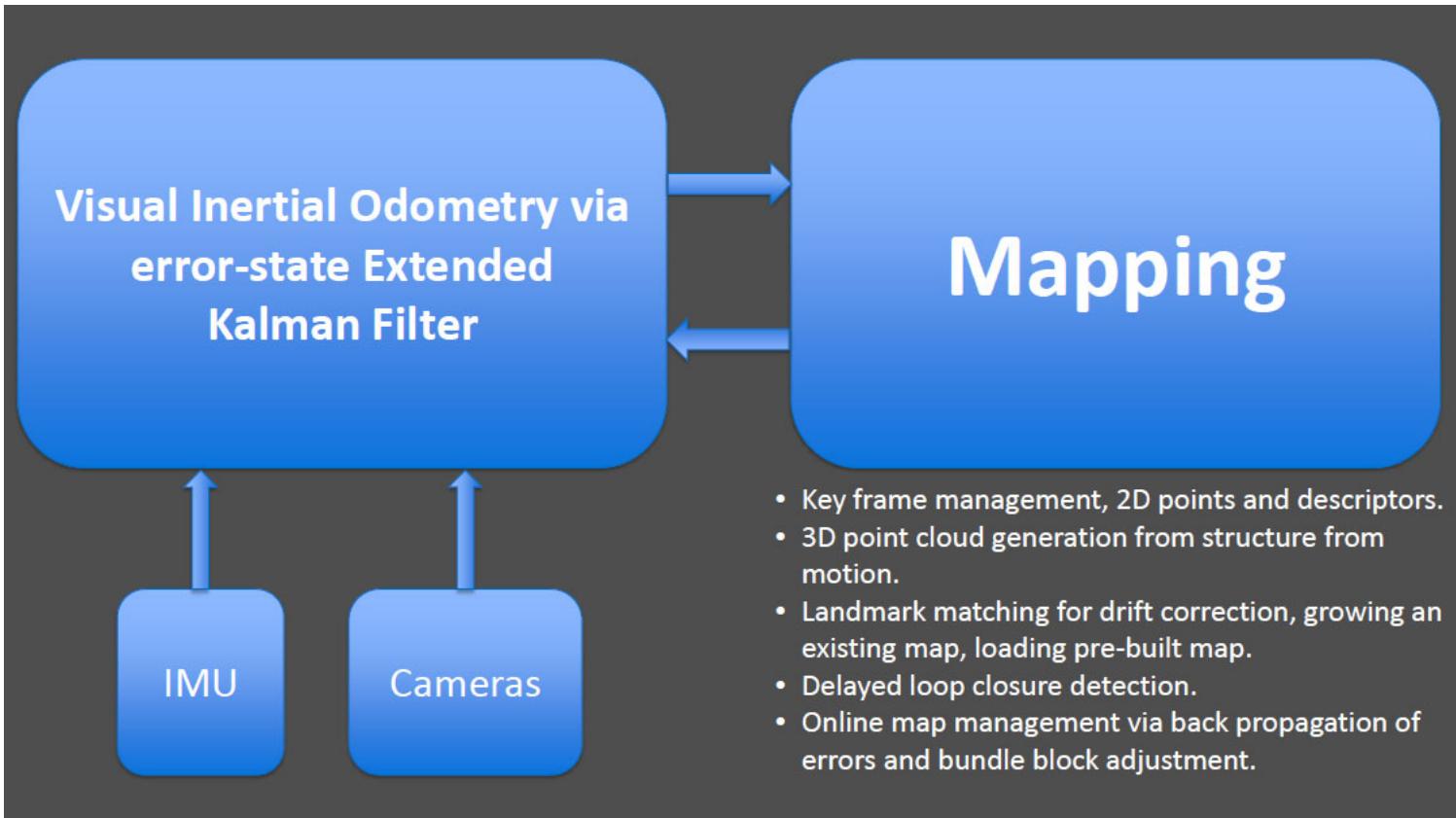
3D Perspective View of Dynamic Map



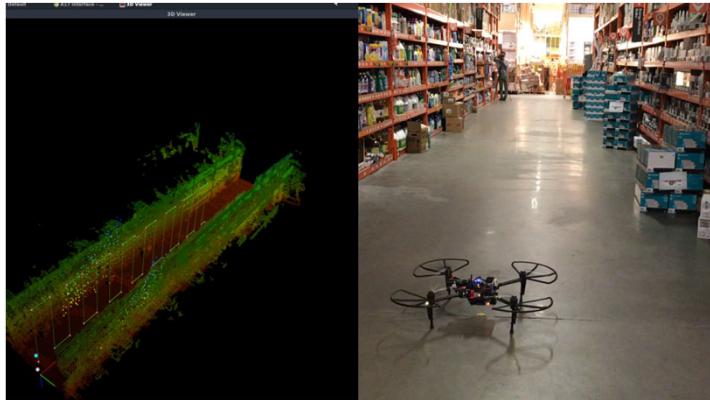
Tightly Coupled Visual-Inertial Odometry



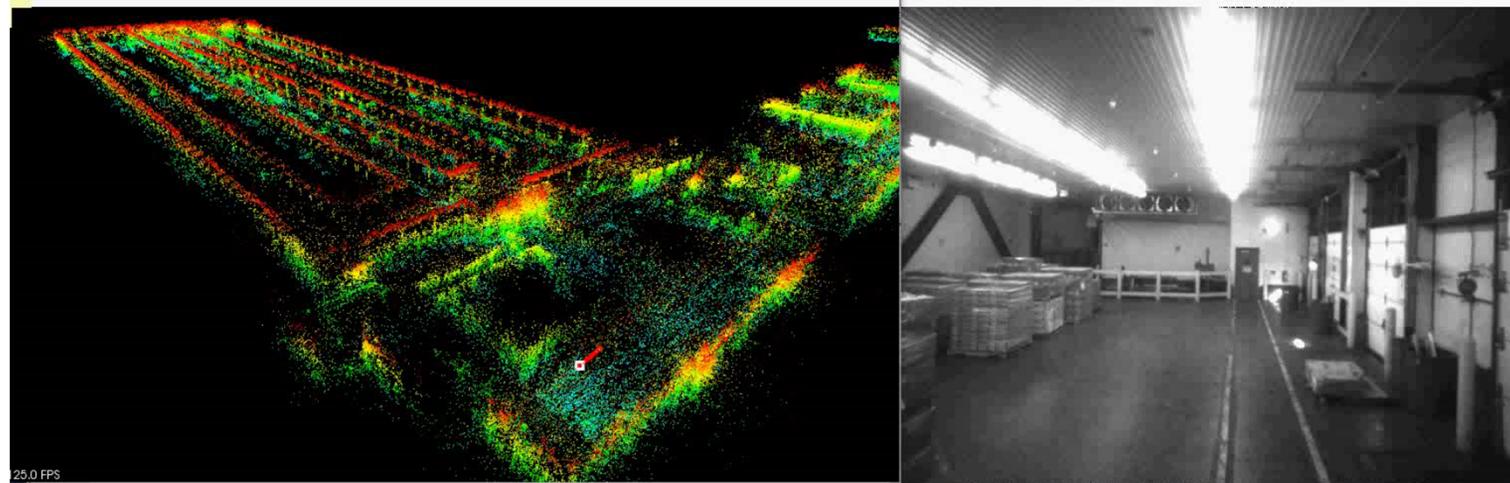
CamSLAM High Level Architecture



Commercial applications based on SRI navigation



ReSCAN



Lineage and
Guide-Robotics

Robotic Applications

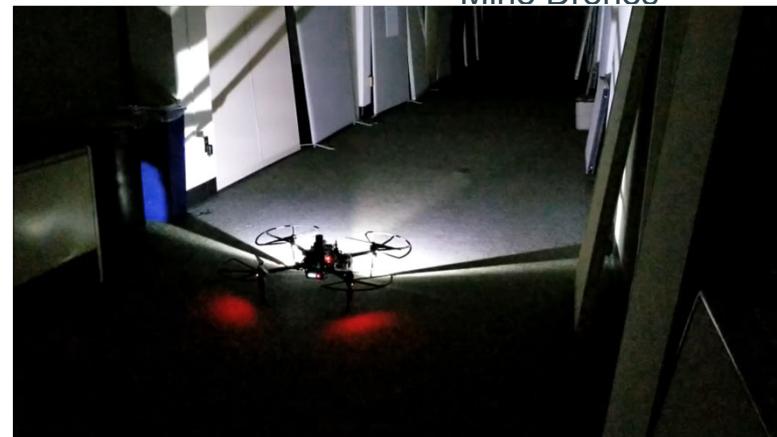
Autonomous Cars



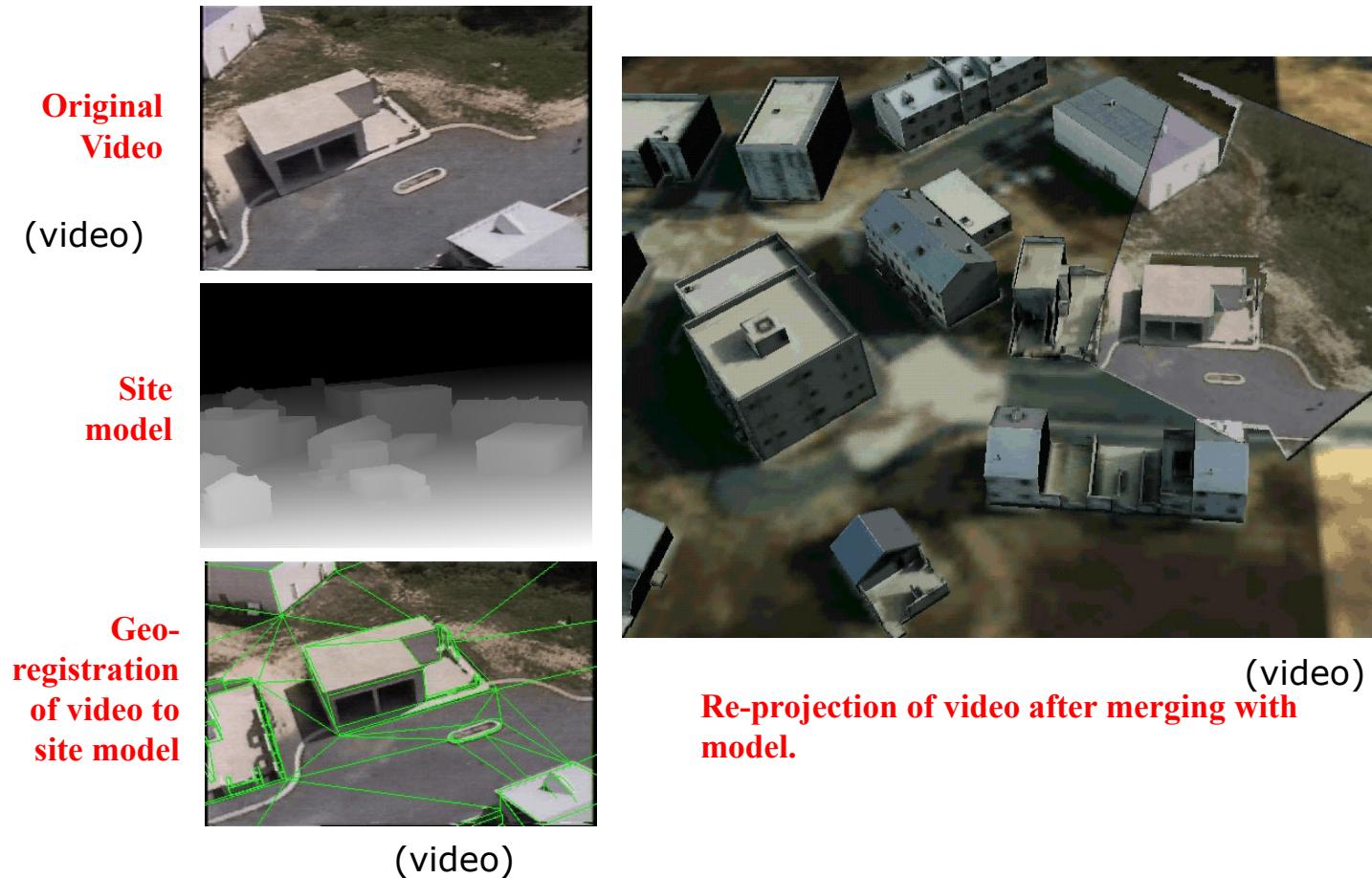
Construction Site Automation



Robotic Followers



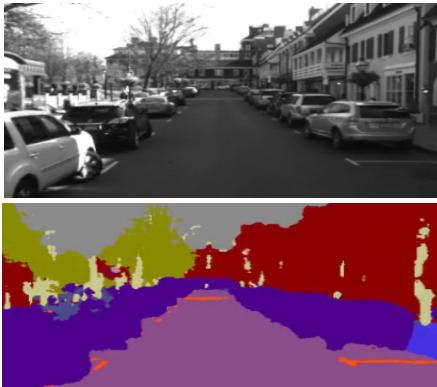
Geo-registration of video to site model ...



New Insight: Semantic Autonomy

A hybrid approach: Develop and integrate both semantic-inference and metric-inference

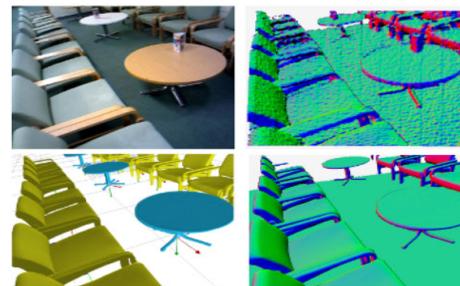
Accuracy



High-level objects are more robust to scene changes, and can be matched across time/space/platforms.

Semantic Autonomy

Efficiency



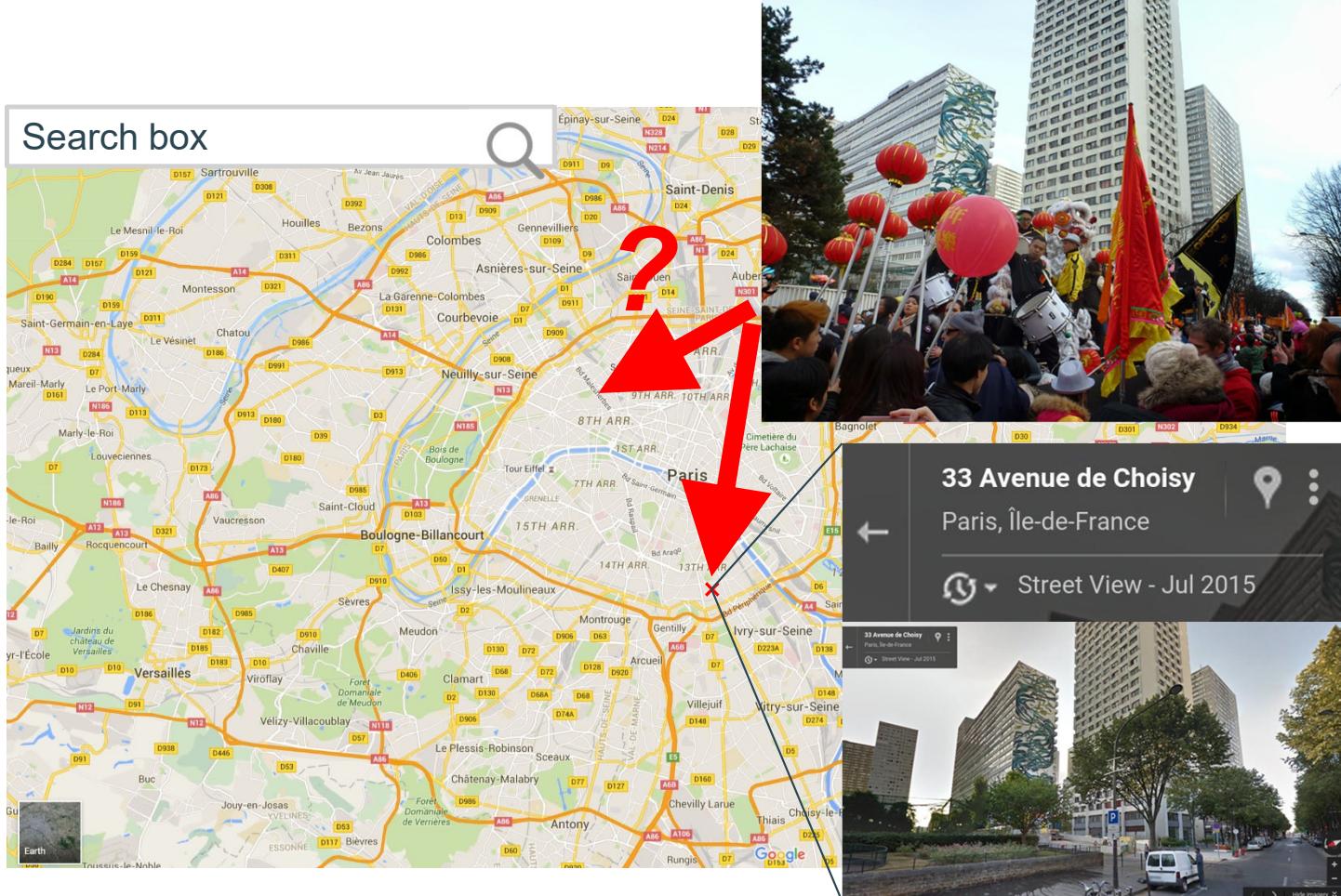
Sharing semantic information reduces bandwidth required for collaboration/ storage.

Applicability



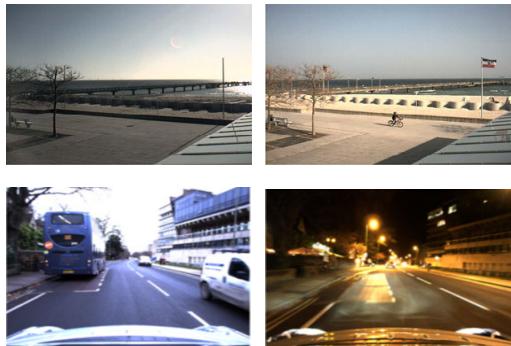
It enables semantic scene representation and natural human-machine interaction for more applications.

Image based Geo-Localization



Cross-Time, Cross-View, and Cross-Modal Geo-registration of ground imagery to reference

Cross-Time



Sample Pairs (Ground RGB)

Cross-View



Sample Pairs (Ground-Aerial RGB)

Cross-Modal



Sample Pairs (Ground RGB-OpenStreetMap)

Low

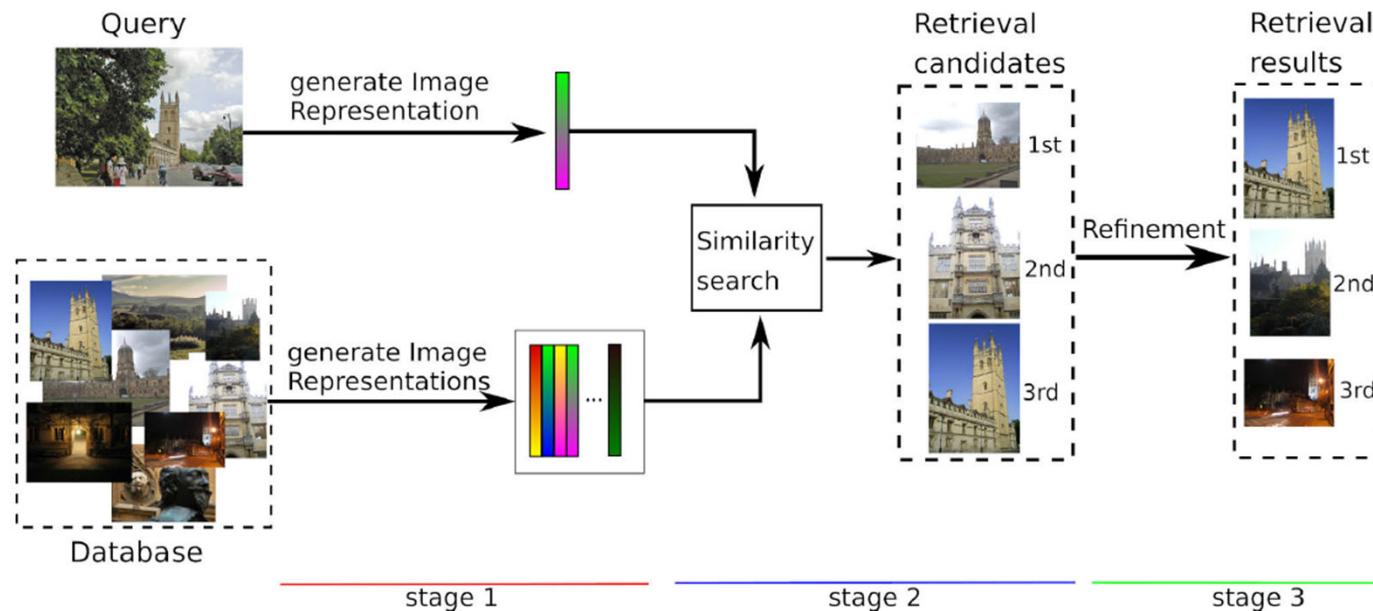
Availability of Geo-Referenced Database

High



Difficulty in Image-Based Visual Localization based on Reference Data

Cross Time Matching: Feature Representation for Image Retrieval



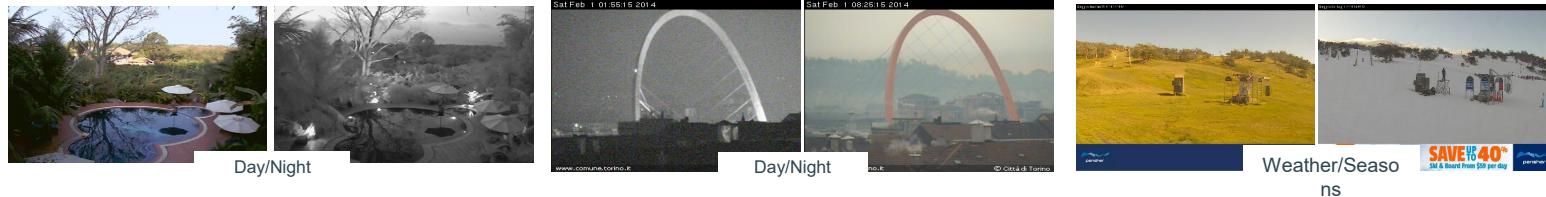
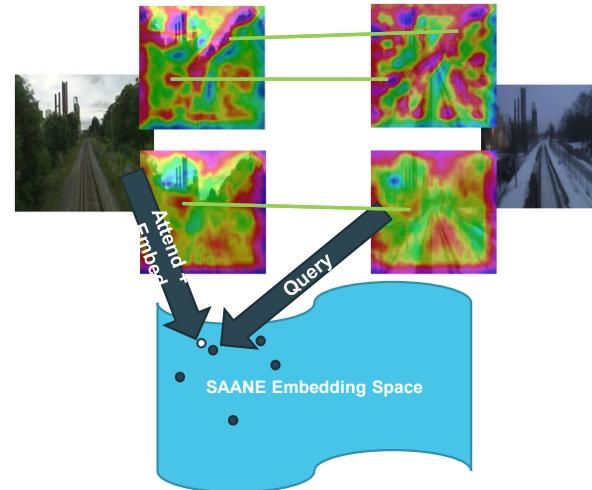
Visual place recognition is commonly formulated as an image retrieval problem. The known places are collected in a database and a new image to be localized is called query. The place retrieval is performed in three logical stages.

- 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.

Image-to-Image Visual Localization

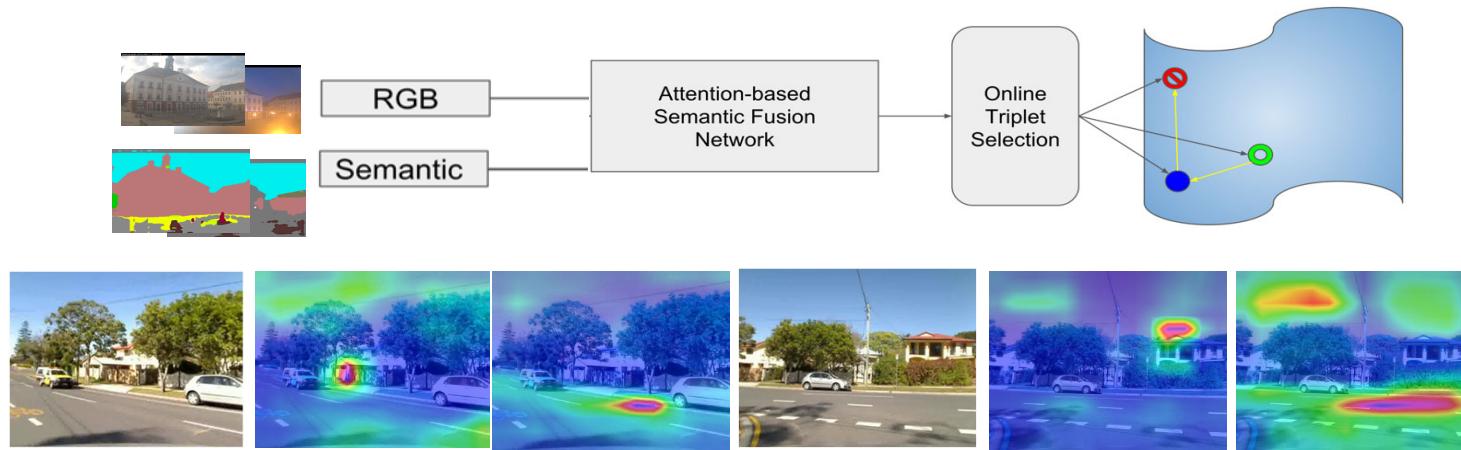
- We propose to use embedding for this problem: A deep-learned compact Euclidean space where distances directly correspond to a measure of data similarity.
- Training data: ~2 million images collected from 2,685 static webcams.



Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
<https://arxiv.org/abs/1812.03402>

Innovation: Attention-based Semantic-Aware Embedding

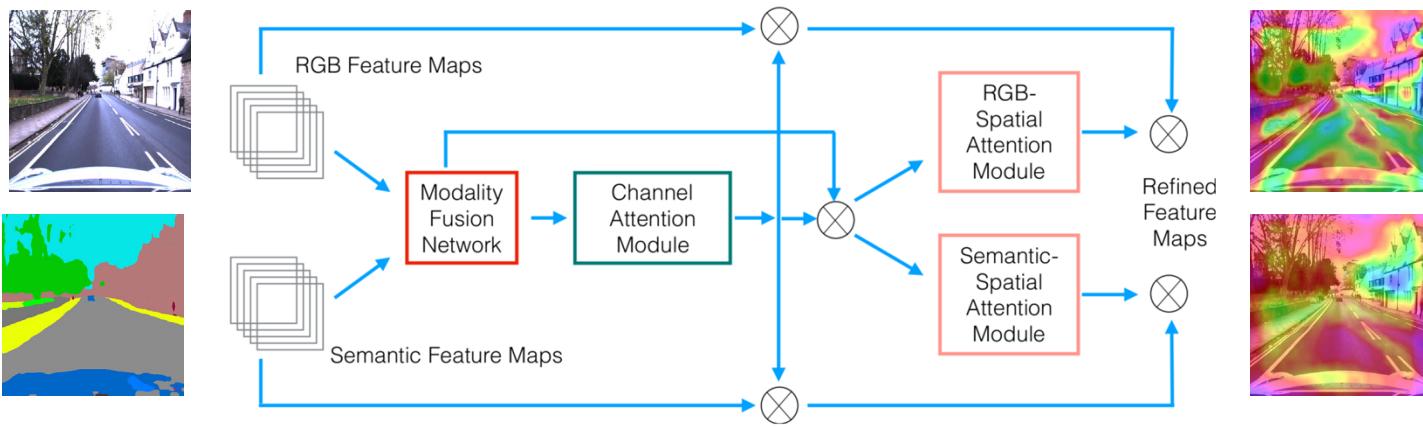
- **Semantic-Aware:** The model incorporates pixel-wise semantic features in learning the image embeddings.
- **Attention-Based:** We train self-attention modules to encourage the model to focus on semantically meaningful spatial regions.
- We evaluate two ways to train attention module: (1) individual attention: on RGB and semantic cue separately, (2) combined attention: on fused feature maps from RGB and semantic cue.



Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
<https://arxiv.org/abs/1812.03402>

Innovation: Attention-Based

- We train a novel formulation of the Convolutional Block Attention Module to encourage our model to focus on semantically-consistent spatial regions.
- The attention network operates in two steps:
 - First, a single attention map is computed for the fused (appearance + semantic) features across the channel dimension to due an initial, multimodal refinement.
 - Second, separate appearance and semantic spatial attention maps are computed to produce the final, refined feature maps.
- Our attention module combined with semantics gives an additional 4% absolute improvement on average.

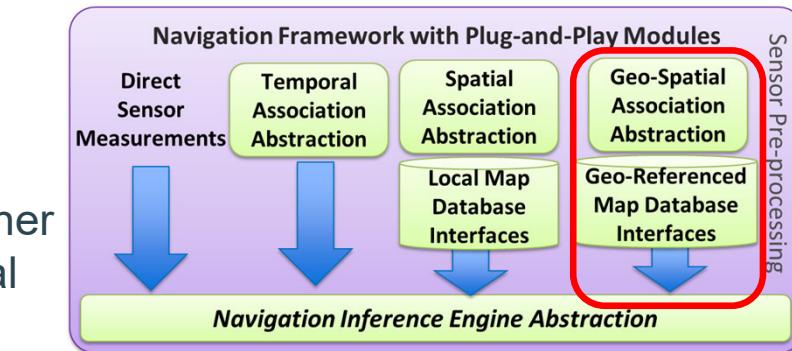


Sanghyun Woo, et al. "CBAM: Convolutional block attention module." ECCV. 2018.

Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
<https://arxiv.org/abs/1812.03402>

Geo-Spatial Association - Day & Night

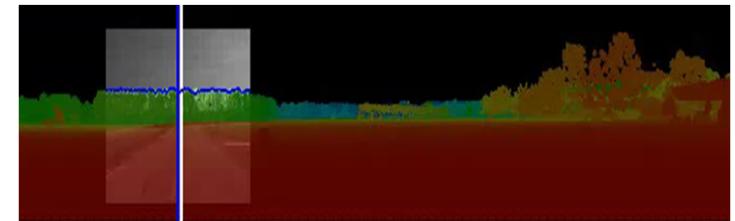
- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.



Cross-view: Global Pose Estimation for ground platforms/ robots

- **Matching to Rendered 3D Overhead Geo-reference Aerial Data**

- Requires Geo-Referenced 3D data, which is often difficult to obtain.



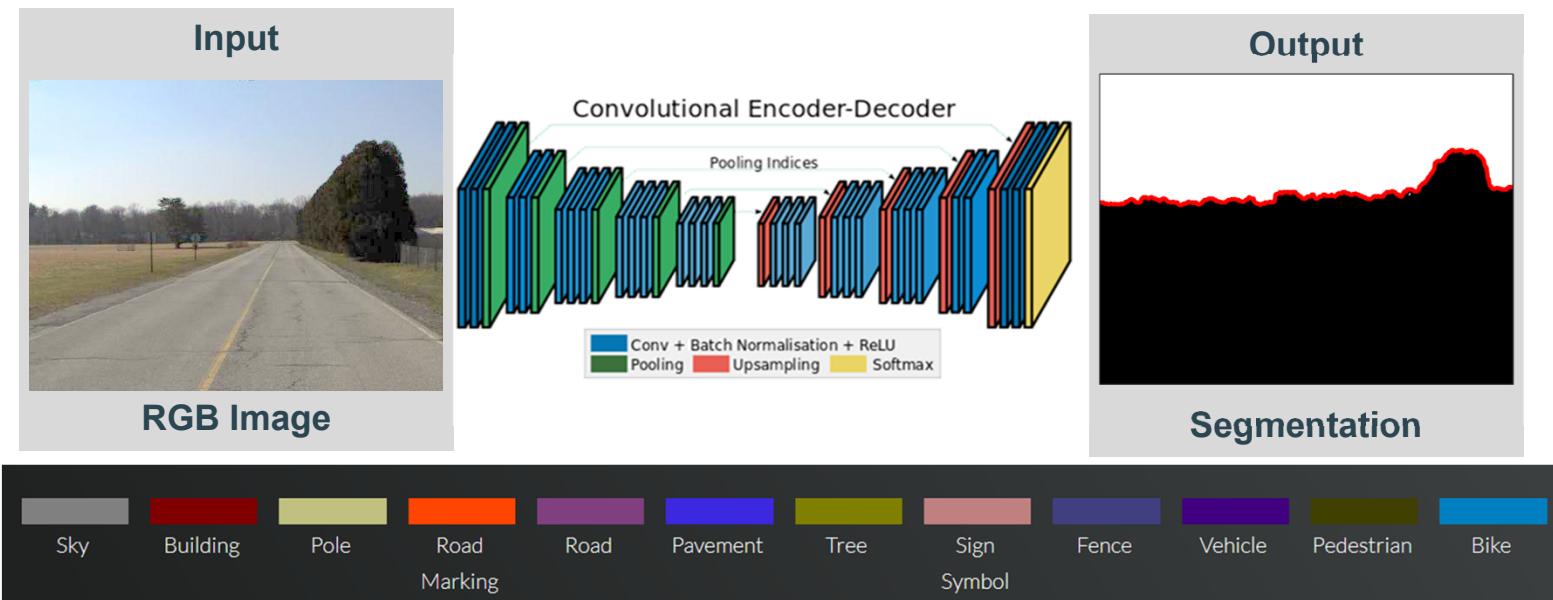
- **Matching to 2D Overhead Geo-reference Data**

- Aerial Satellite Reference: Widely available, Challenging cross-view matching.



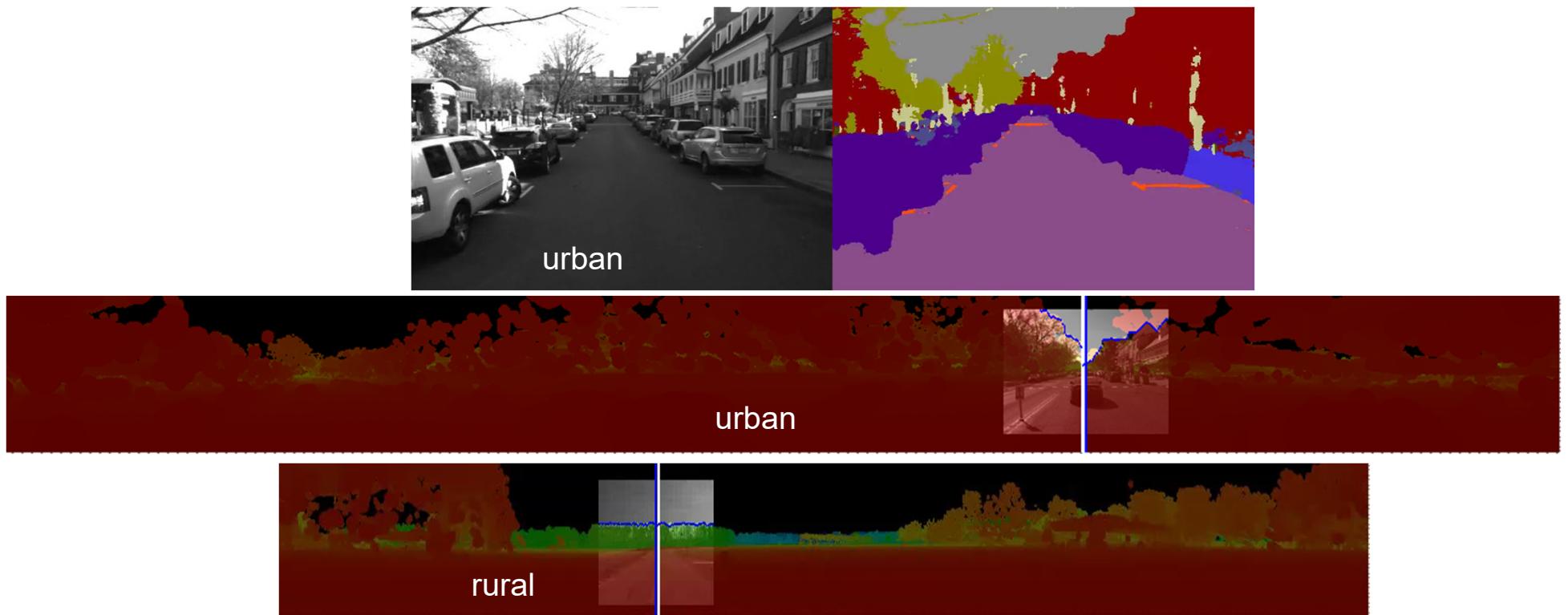
Geo-registration of ground imagery to aerial reference with 3D terrain

- 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.



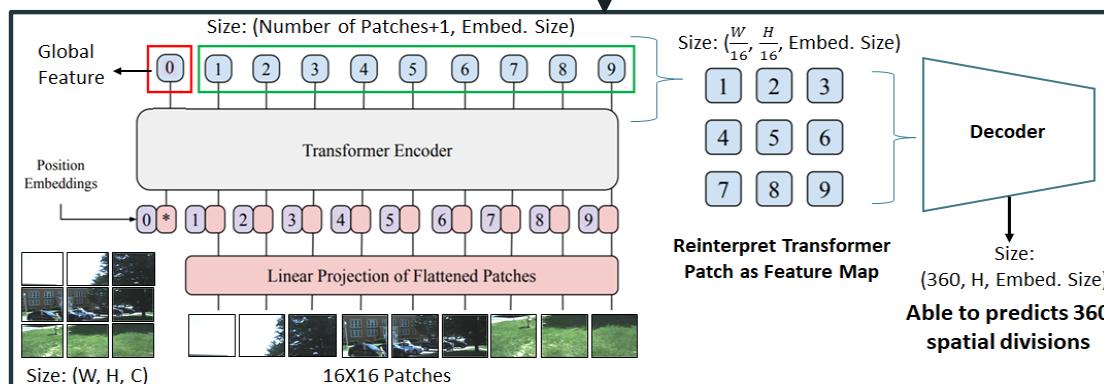
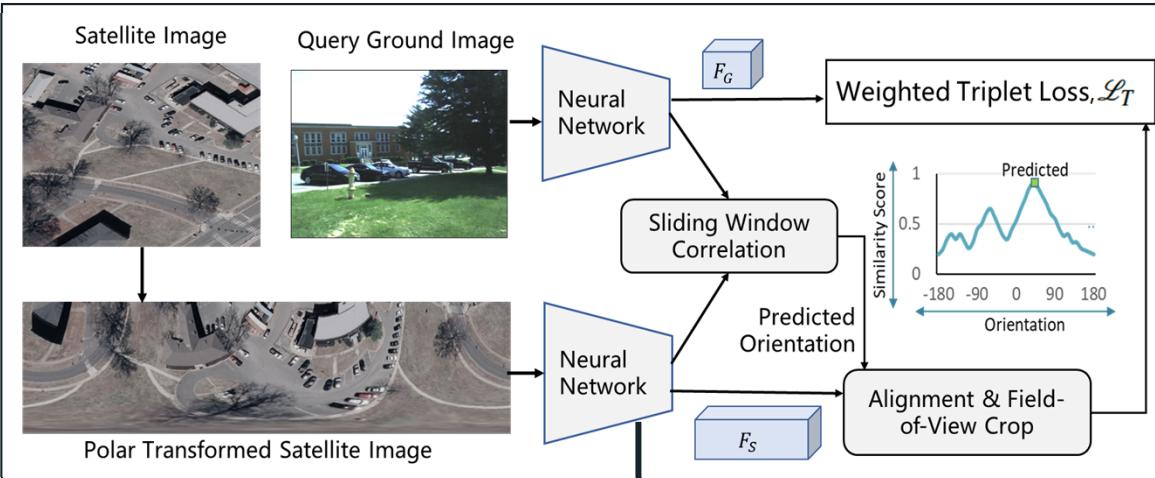
Geo-Spatial Association – Semantic Geo-Registration

We 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.

Geo-Registration of ground imagery 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.

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

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

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

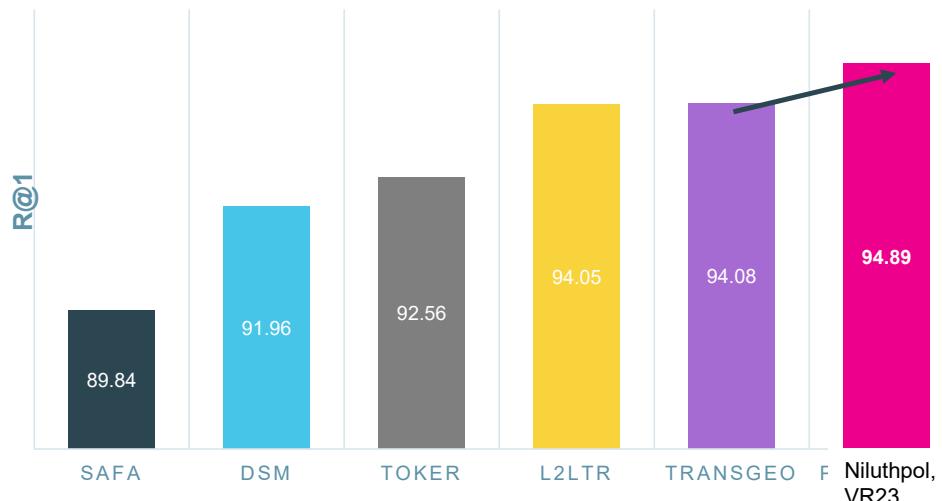
$\mathbb{F}_{\hat{S}}$ is a non-matching satellite image feature embedding for ground image feature embedding \mathbb{F}_G , and \mathbb{F}_S is the matching satellite feature embedding.

\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.

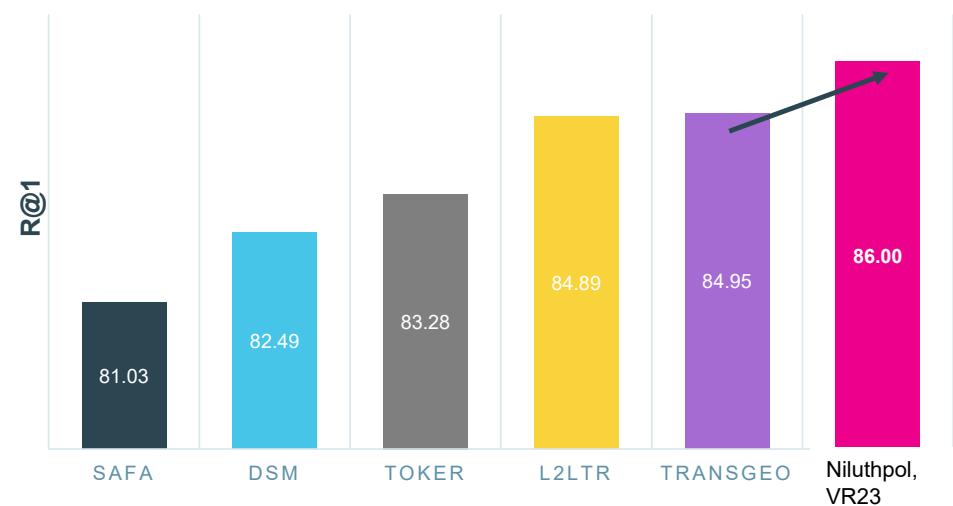
SRI International®

Location Estimation Performance

LOCATION ESTIMATION ON CVUSA



LOCATION ESTIMATION ON CVACT



- 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

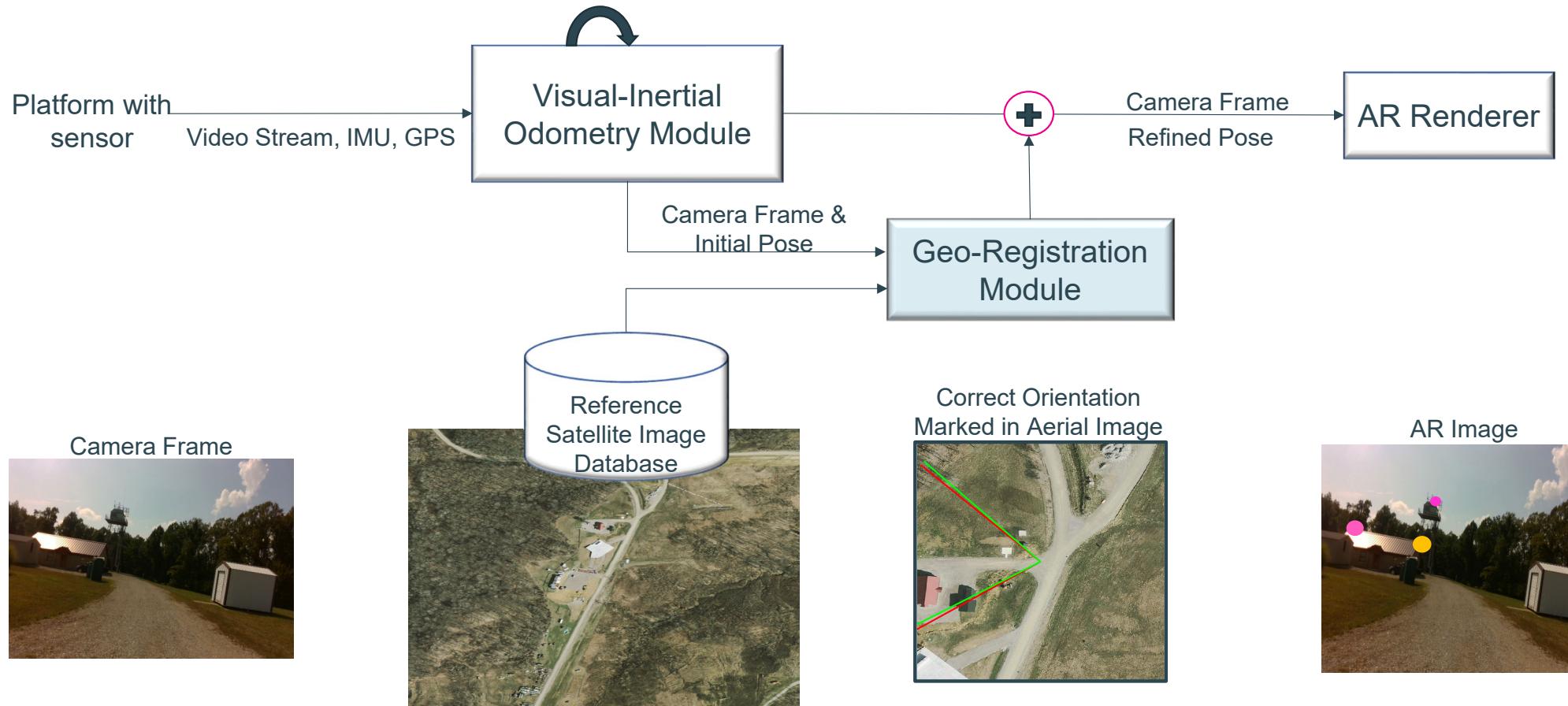
Orientation Estimation on CVUSA

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

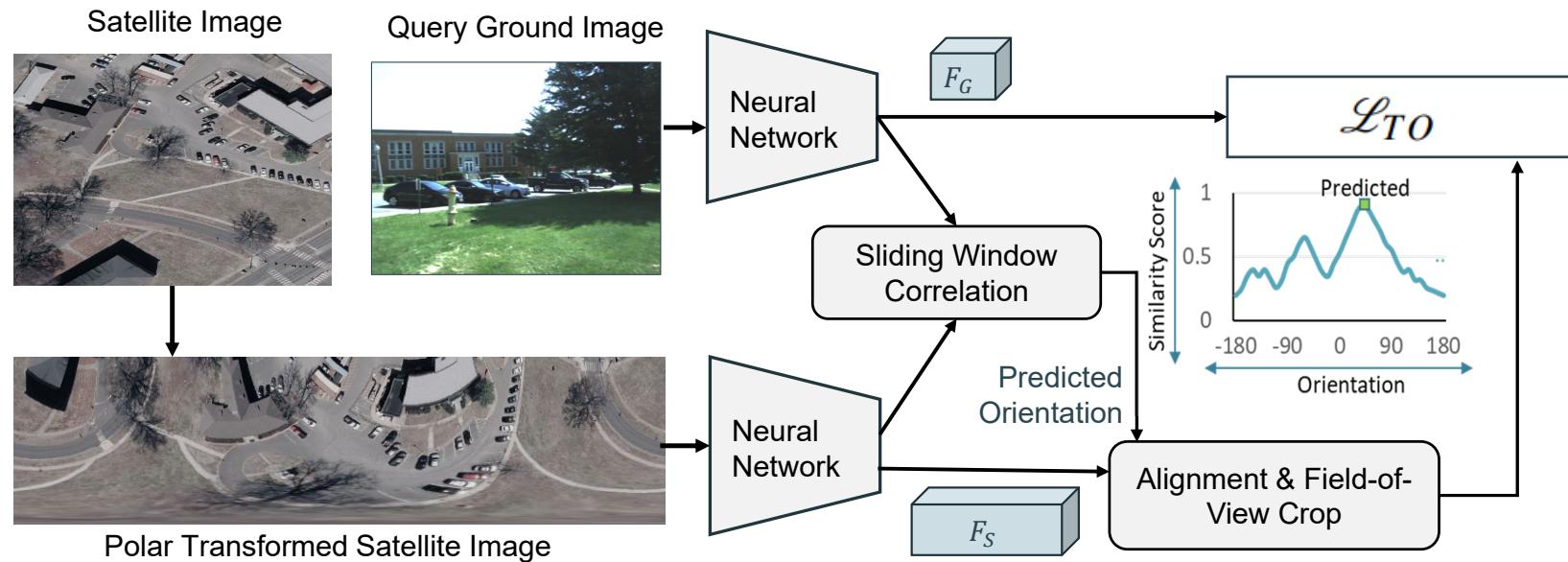
Geo-Alignment for estimating pose



Handling Real Sequences for navigation 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 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

Updated Training to Adapt for Prediction on Real-World Sequences with varying field of view imagery

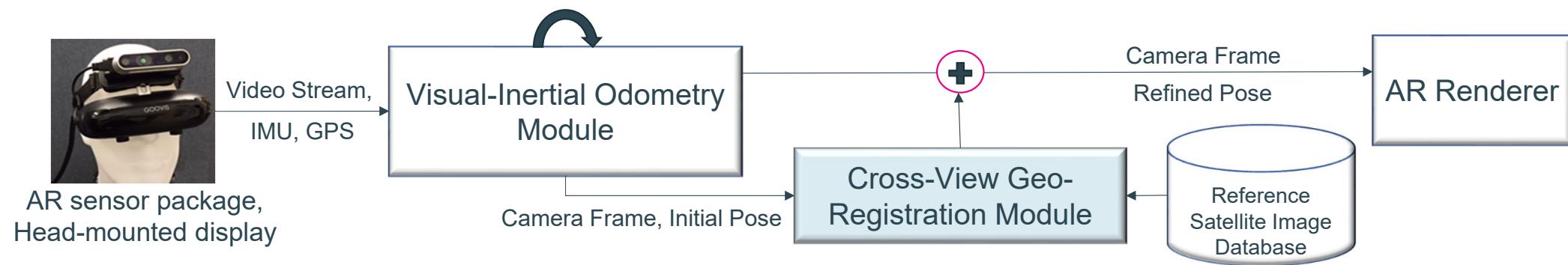


- Train base model using benchmark dataset (e.g., CVUSA) with orientation weighted triplet loss L_T .
 - First with panorama (360 degree) ground images and reference satellite. This helps learn robust features.
 - Next, fine-Tune with limited field-of-view (i.e., 70 degree) ground images. This helps adapt to our system setting.
- Fine-Tune with our collected Real-sense (70 degree) data pairs minimizing L_{TO} with high weights to L_{ORI}
 - High L_{ORI} helps model adapt to orientation estimation and low L_{GS} limits the effect of location-based error.
 - Our images are collected densely. location loss L_{GS} have difficulty contrasting between close location pairs.

Confidence metric and integration of geo-registration with navigation module

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

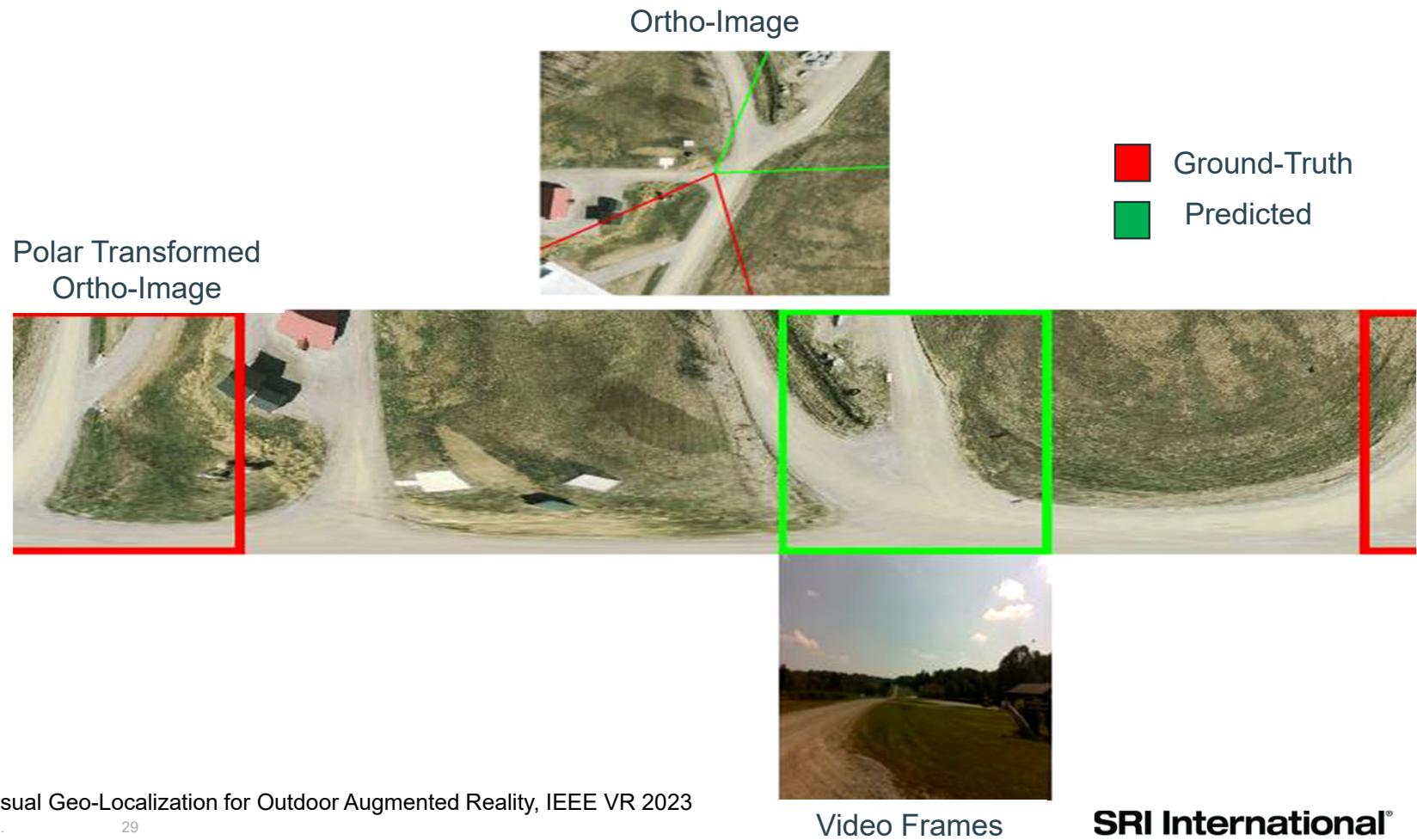
- Works reasonably well with panorama images. However, when the camera FoV is small, a single frame have limited context and the estimate based on a single frame is not reliable/stable for AR/ navigation application.



Approach for continuous Sequences of video frames - 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

Experiments on Real-World Navigation Sequences



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.

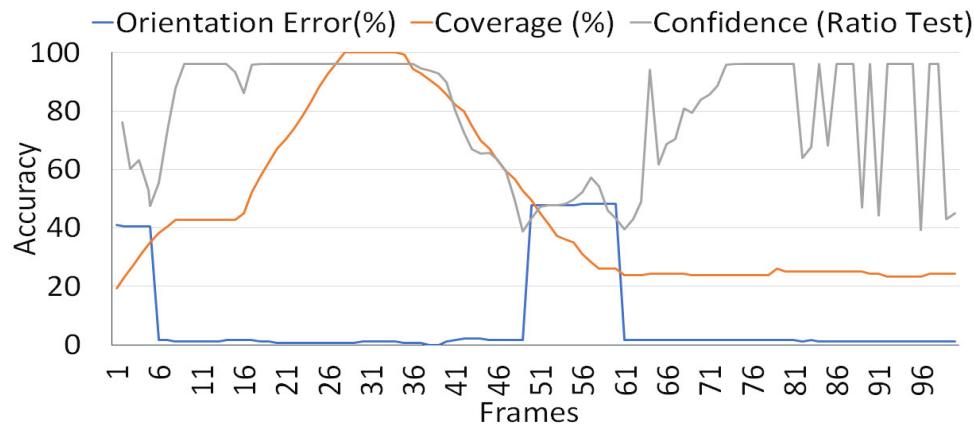


Fig. Orientation Estimation on a 100 frames from Set2, Sequence from Johnson County, IN. Last 20 frames are used in estimation.

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

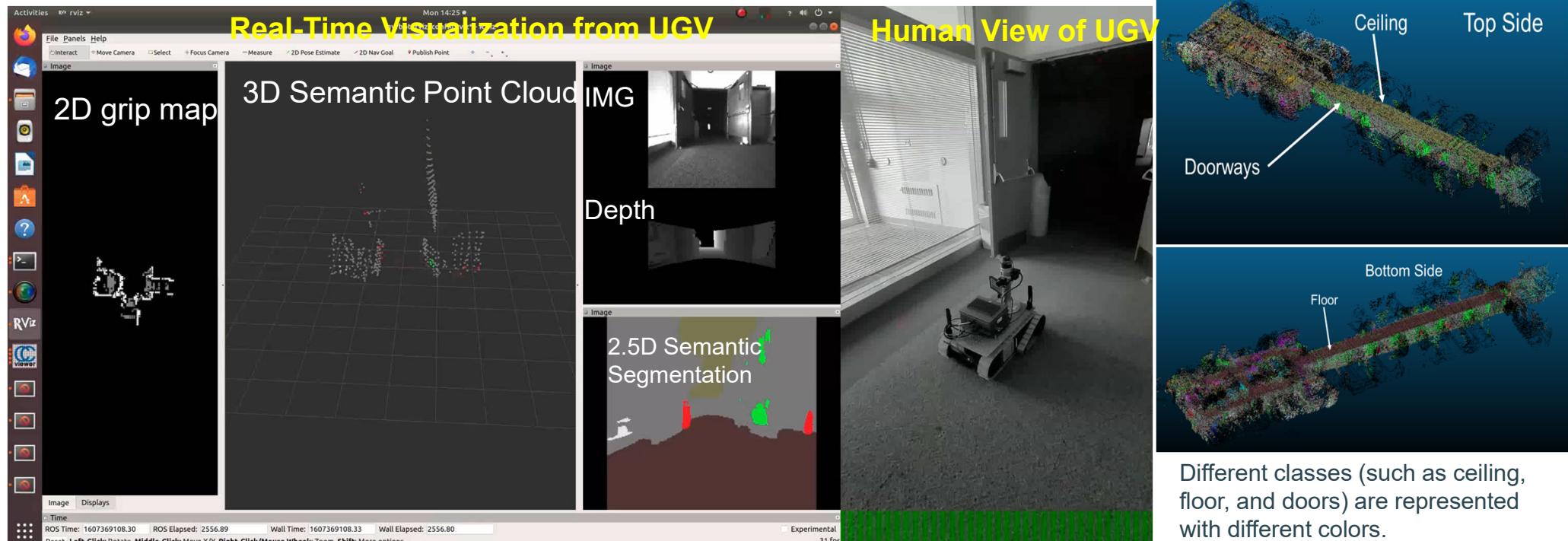


- Comparable results even in GPS and Magnetometer denied case.

Video

Real-Time Semantic Scene Understanding for Robotic Autonomy

- SRI developed the first real-time semantic navigation and mapping system for robots to operate in new unknown environments, under both normal lighting and visually-degraded conditions.



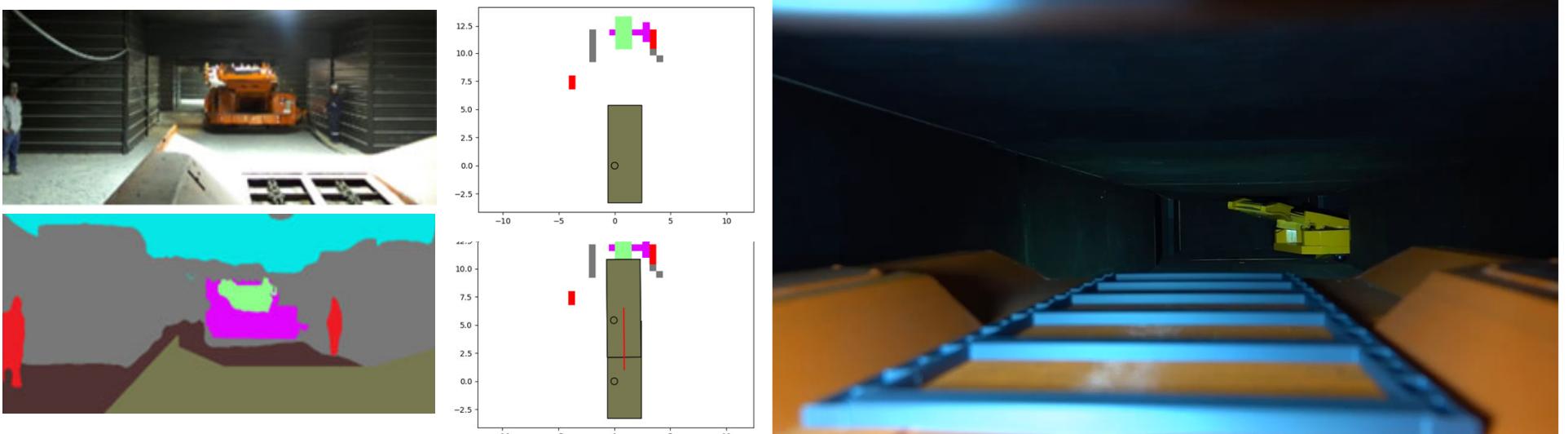
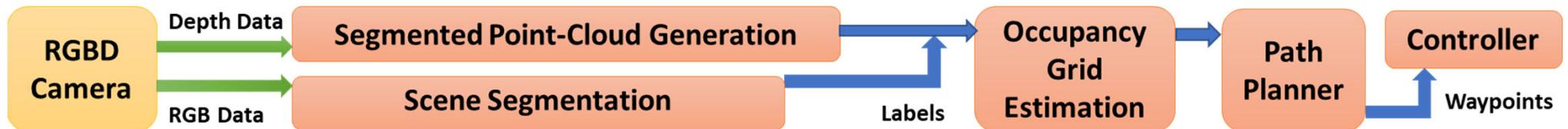
Han-Pang Chiu et al, **SIGNAV: Semantically-Informed GPS-Denied Navigation and Mapping in Visually-Degraded Environments**. Winter Conference of Computer Vision (WACV), 2022..

Han-Pang Chiu et al, **Striking the Right Balance: Recall Loss for Semantic Segmentation**. International Conference on Robotics and Automation (ICRA), 2022..

© 2023 SRI International. All Rights Reserved.

Semantic Autonomy in Challenging Environments

Project: NIOSH Autonomous Docking – Provide autonomous navigation solutions capable of navigating a vehicle based on semantic scene information in challenging environments such as underground mines.



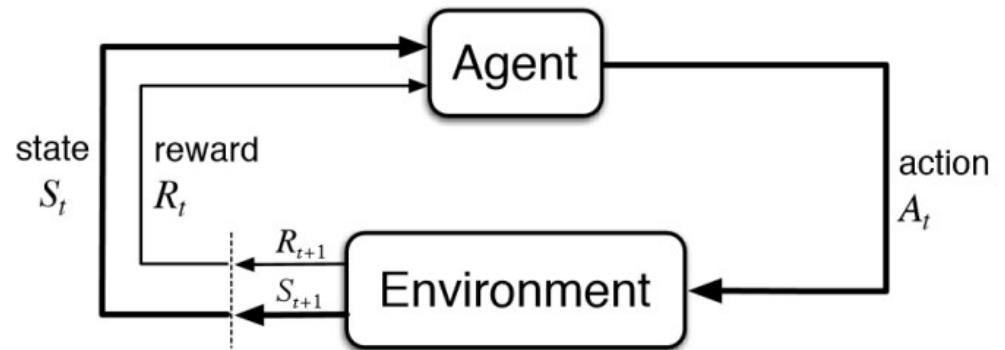
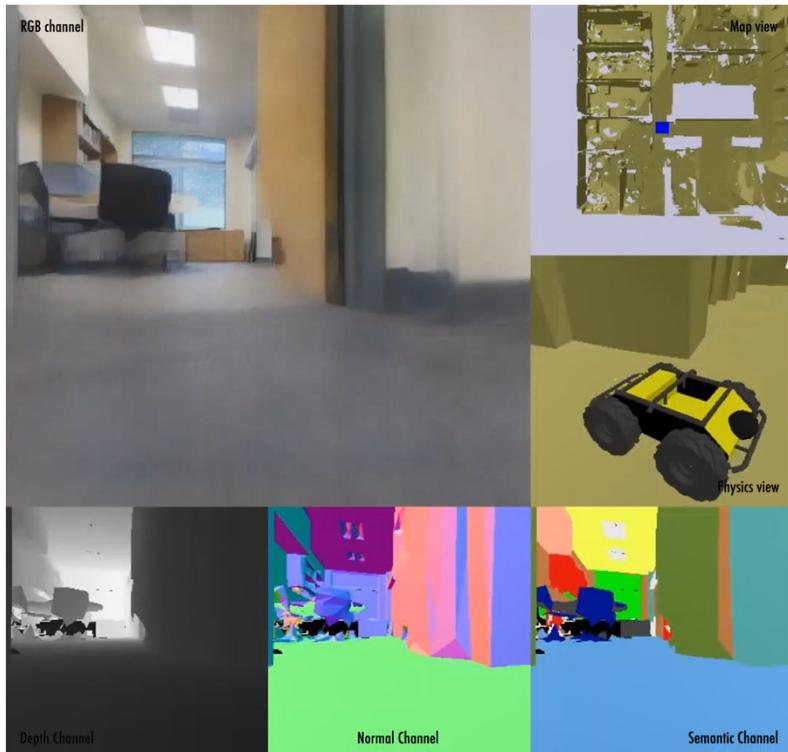
J. Sottile, A. Rajvanshi, Z. Agioutantis, A. Krasner, S. Schafrik, J. Rose, M. Sizintsev, H. Chiu, Evaluating the Efficacy of Autonomous Shuttle Cars Tramming and Docking Sensor and Control Packages, SME, 2023.

A. Rajvanshi, A. Krasner, M. Sizintsev, H. Chiu, J. Sottile, Z. Agioutantis, S. Schafrik, J. Rose, Autonomous Docking Using Learning-Based Scene Segmentation in Underground Mine Environments, IEEE SSRR, 2022

Traditional Robotic Autonomy: Limitation

- Rely on a highly accurate metric map of low-level features features that is built beforehand or constructed using simultaneous localization and mapping (SLAM) algorithms during the mission.
 - Scene appearance can change over time. The map size can be large.
- Path planning will fail if no reliable map or absolute position can be used.
- No accumulation of experience

AI-Enabled Autonomy



Deep Reinforcement Learning (DRL): Learn how to achieve a goal through a direct mapping from situations (sensor readings) to actions (control commands), by trial-and-error interactions with its environment.

- **Better generalization to new environments:** Navigate in a new spaces without having seen the same type of place before; i.e., no underlying model of “known” place types.

Traditional vs AI Methods for Robotic Autonomy

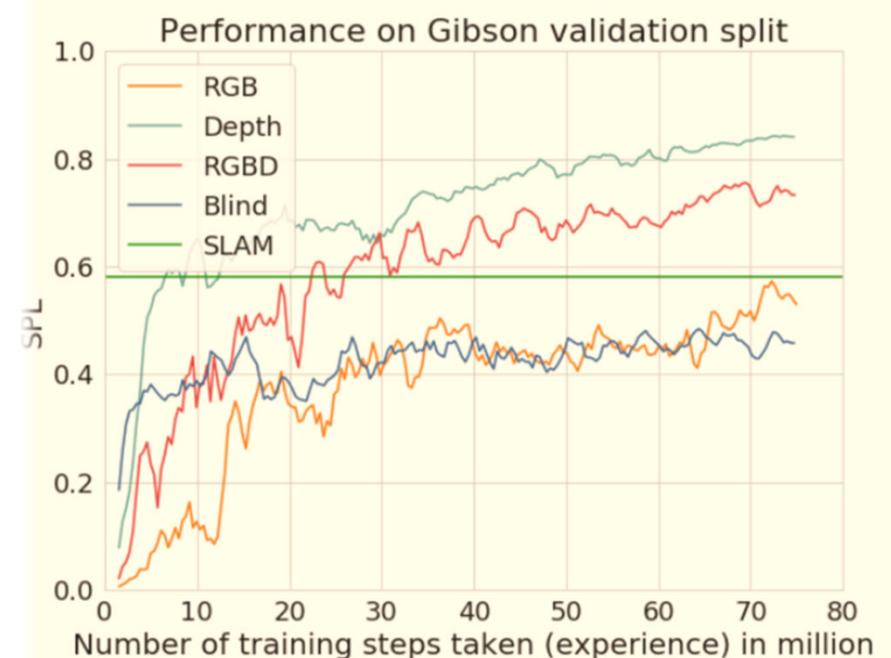
- DRL is better than traditional methods, given enough experience.

- **Require lots of training data**

- 2.5 billion steps of trainings
 - Equal to 80 years of human experience

- **Inherent non-interpretability:**

- Create “black boxes” for reasoning
 - Lack capabilities to explain or reason its behaviors or actions, which are required to interact with humans.

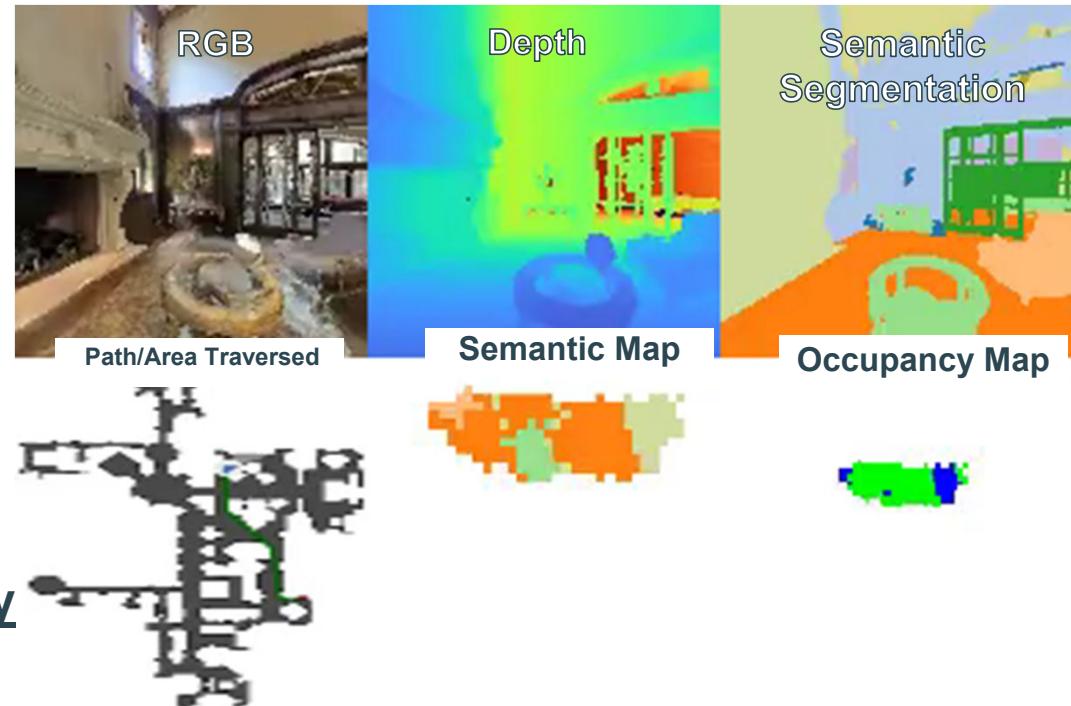


D. Mishkin et al., "Benchmarking Classic and Learned Navigation in Complex 3D Environments." *arXiv preprint arXiv:1901.10915* (2019).

Semantic Reasoning for AI-Enabled Autonomy

- Employ semantic scene structures to reason about the world and pay particular attention to relevant semantic landmarks to develop navigation strategies.
- Enhance DRL with **explicit semantic scene information as maps or graphs** to learn more efficiently from reduced training data.
- Preserve explicit geometric relationships within semantic information to improve the performance.

Achieved state of the art accuracy with less than 20% training data.

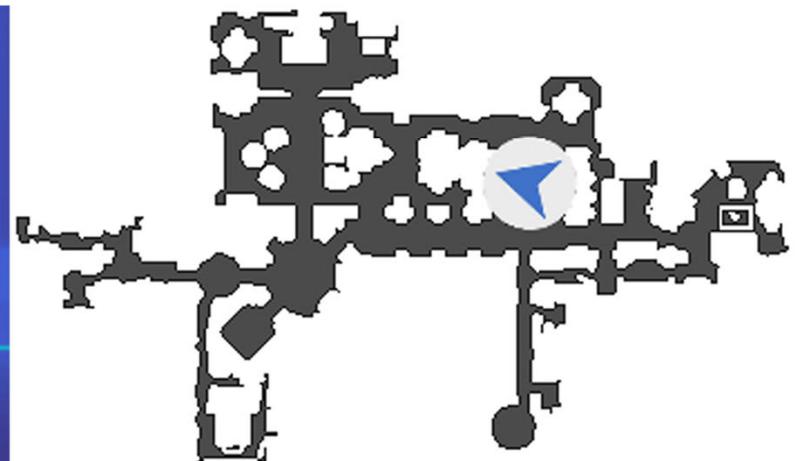
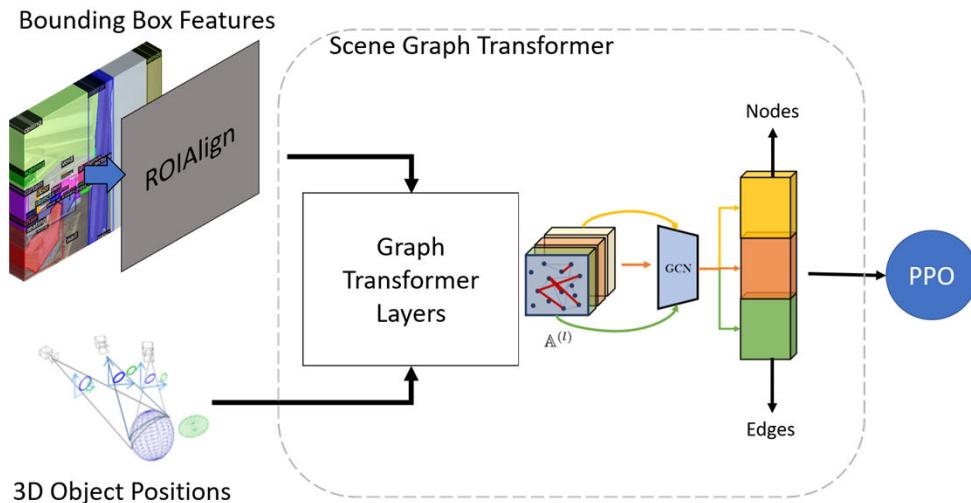
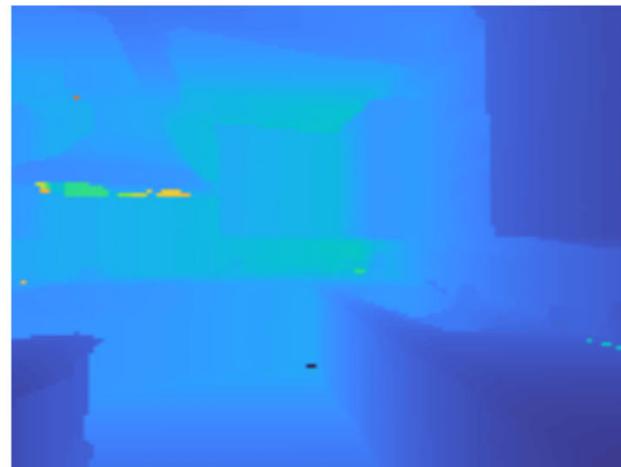


Han-Pang Chiu et al., **MaAST: Map Attention with Semantic Transformers for Efficient Visual Navigation**. IEEE International Conference on Robotics Automation (ICRA), 2021.

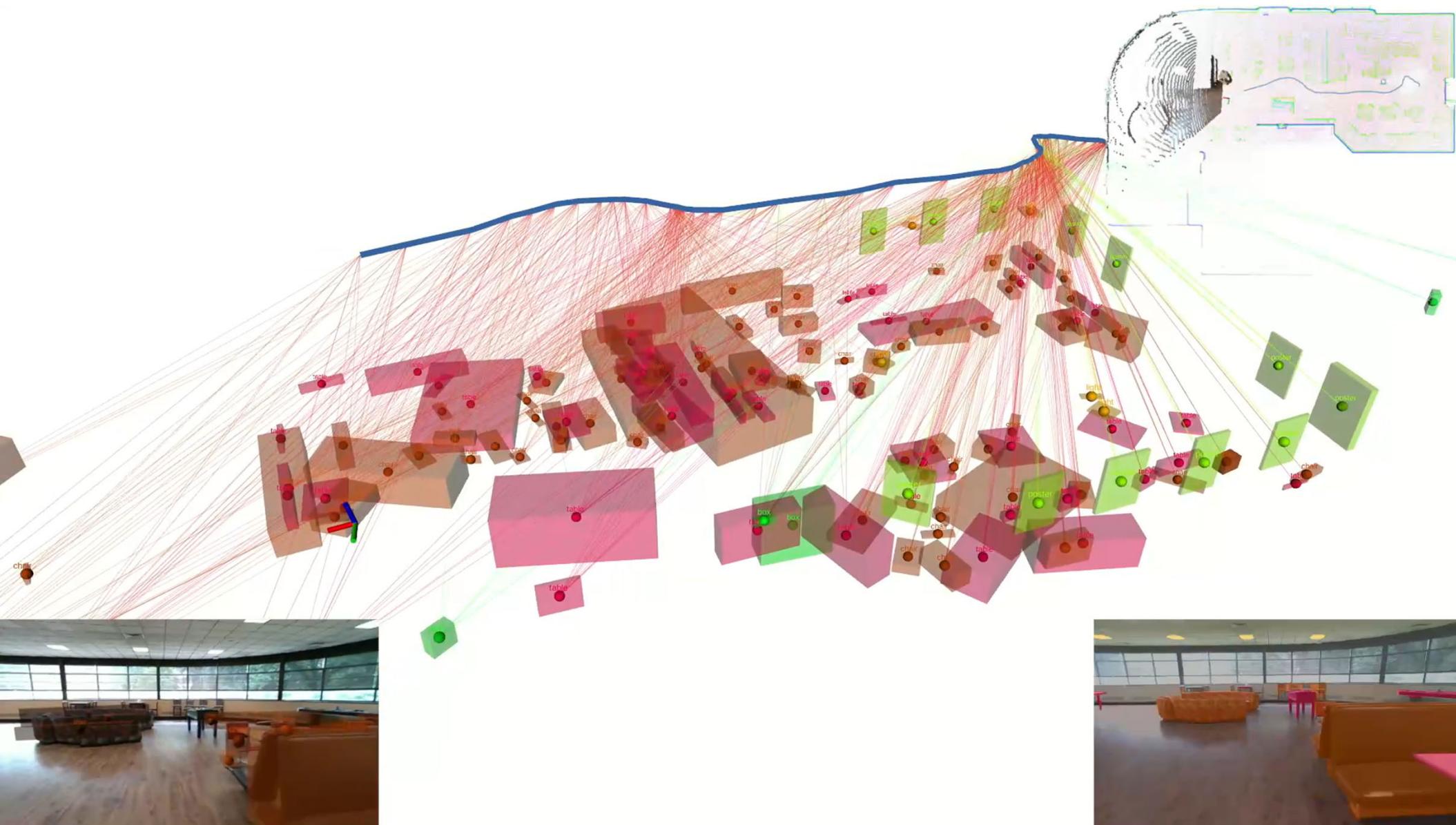
Han-Pang Chiu et al., **SASRA: Semantically-Aware Spatio-Temporal Reasoning Agent for Vision-and-Language Navigation in Continuous Environments**. International Conference on Pattern Recognition (ICPR), 2022.

Exploration by Building Scene Graphs

- The agent learns to explore/navigate by predicting the scene graph of the environment by accumulating from frame-to-frame
- The agent simultaneously learns to navigate and to construct a better, more generalizable scene representation to be passed up the autonomy stack for higher-level planning.

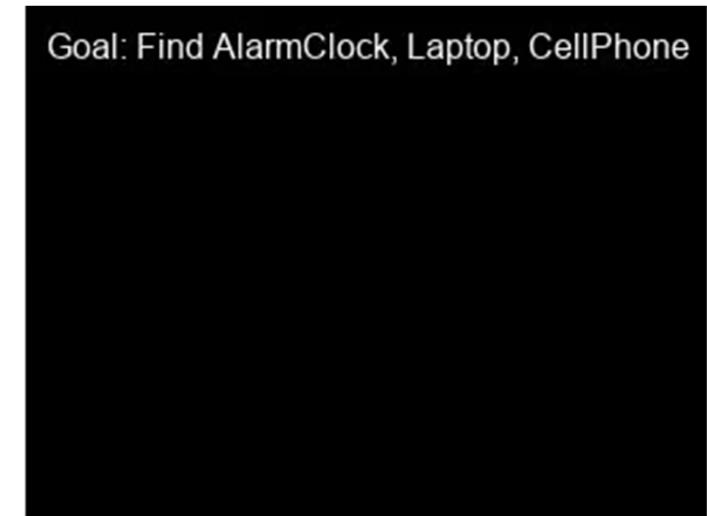


Han-Pang Chiu et al., **GraphMapper: Efficient Visual Navigation by Scene Graph Generation**. ICPR 2022.



SayNav: Grounding Large Language Models for Dynamic Planning to Navigation in New Environments

- Use scene graphs to ground LLMs for navigation tasks in unknown large-scale environments
 - Reduce the learning complexity by using a two-level planning architecture - utilizing LLMs to generate a high-level step-by-step plan which can be executed by a pre-trained low-level planner that maps one step into a sequence of primitive actions.



*Han-Pang Chiu et al. "SayNav: Grounding Large Language Models for Dynamic Planning to Navigation in New Environments", arXiv:2309.04077.



Goal: find a person hiding in the area

Questions