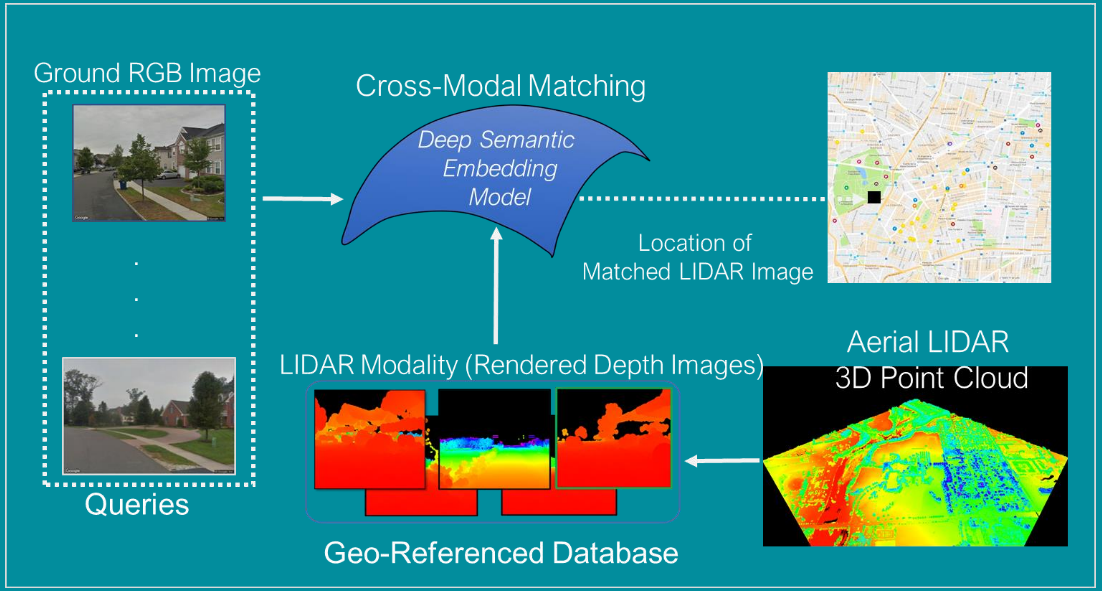
Cross-Modal Geo-Localization: Image-to-3D Coarse Search & Fine Alignment

Han-Pang Chiu

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

June 20, 2021



RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization

SRI International®

Outline

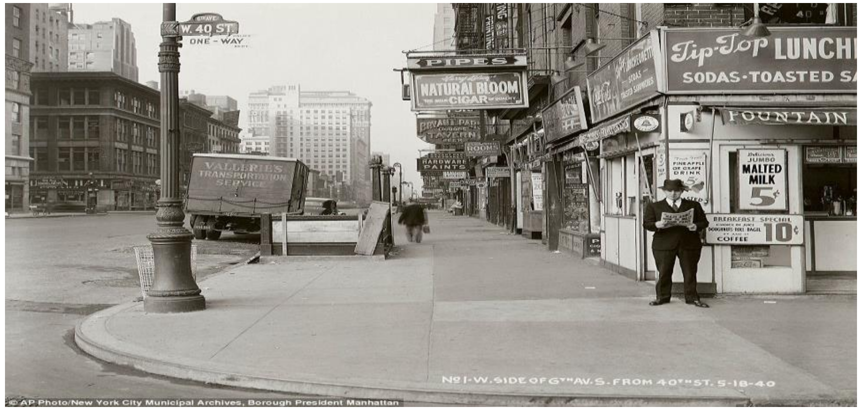
- **Cross-Modal Geo-Localization**
- Coarse Search
- Fine Alignment
- Conclusion
- Q & A

Image-based Geo-Localization

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

Applications

Historical Imagery



Improve GPS Accuracy

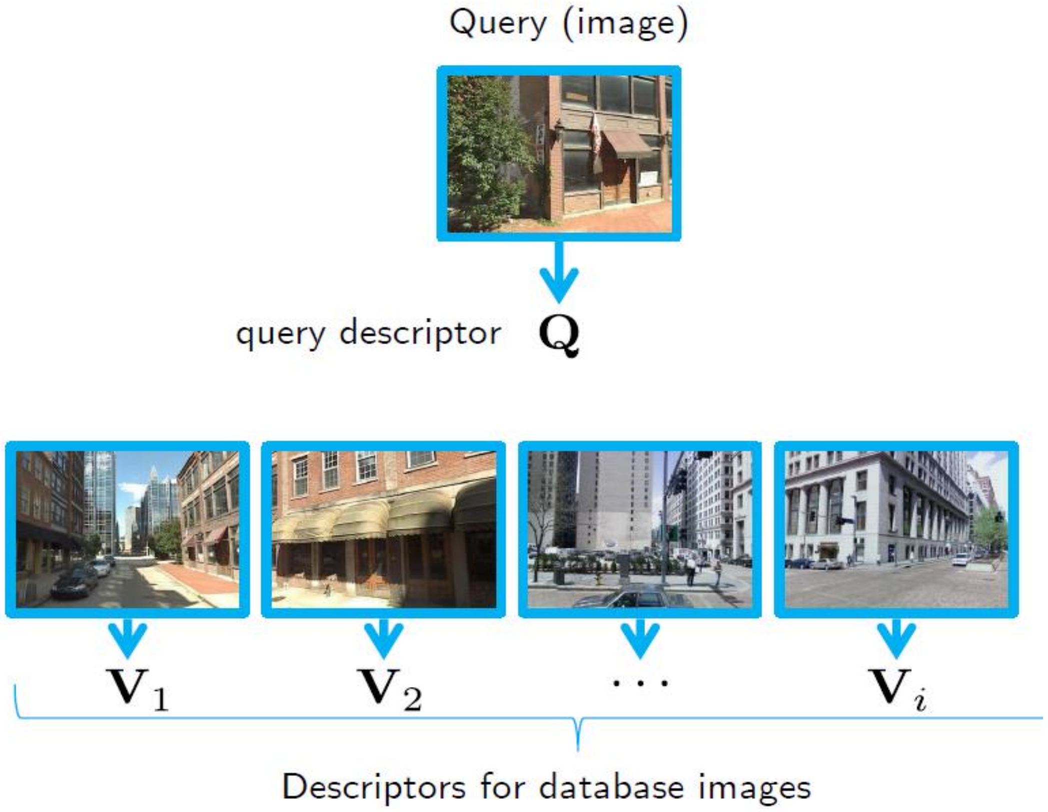
Personal Photo Album

GPS Denied/Challenged Environments

SRI International

Image-Based Visual Geo-Localization – Coarse Search

Matching a query image to a geo-referenced databases (Database Retrieval), which is also called Place Recognition in some fields.



Geopose $_1$ Geopose $_2$ . . . Geopose $_i$

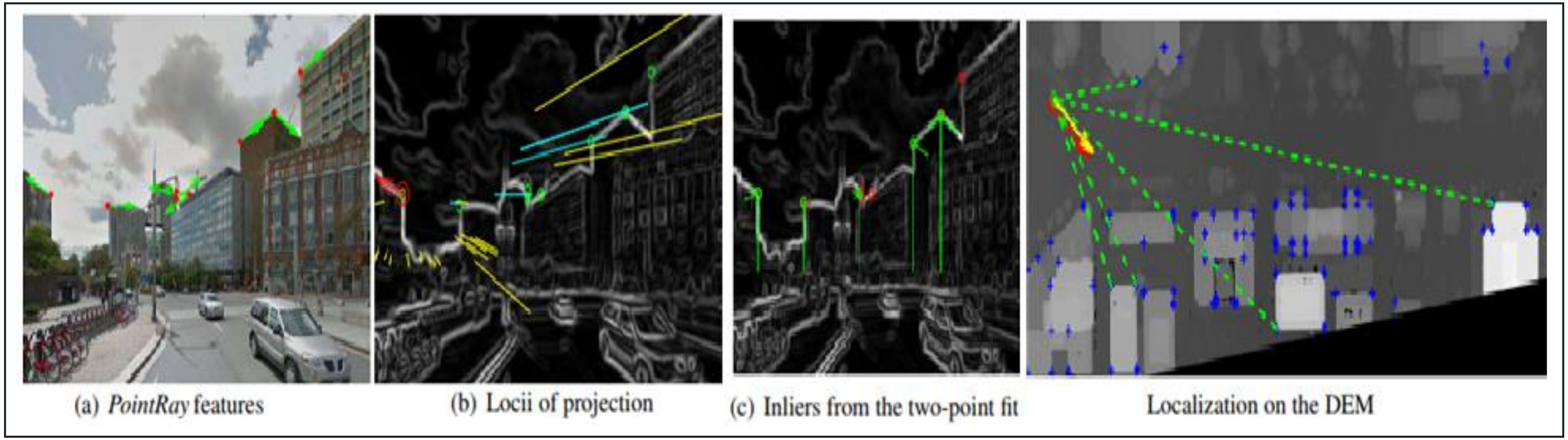
$$i^\star = \underset{i \in 1 \dots N}{\arg\max} \; s(\mathbf{Q}, \mathbf{V}_i)$$

Nearest Neighbor search

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.

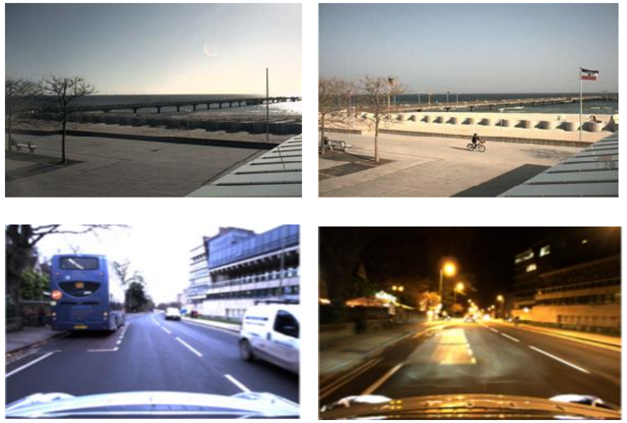
- It requires detailed 3D information in the database (such as 3D point cloud).
- It is also called geo-registration.



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

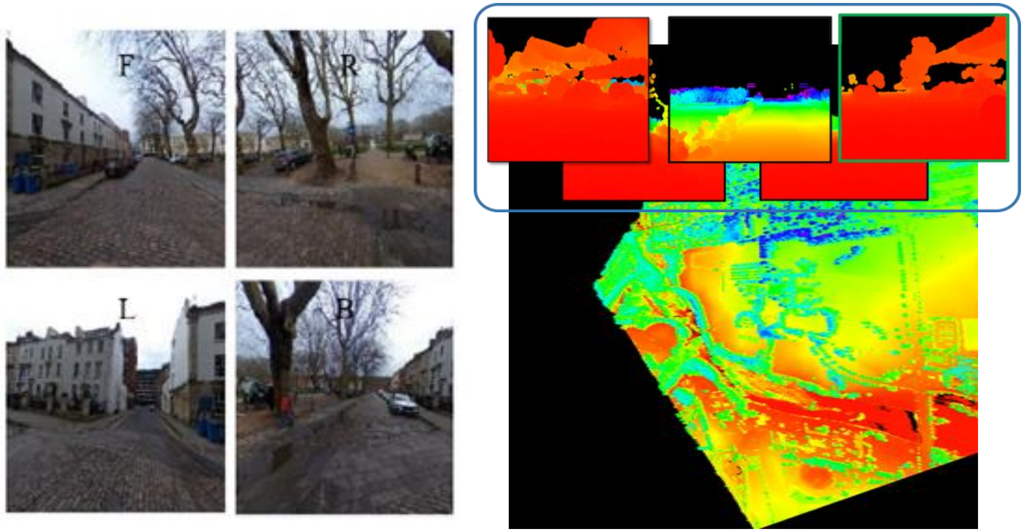
Difficulty in Image-Based Visual Localization

Outline

- Cross-Modal Geo-Localization
- **Coarse Search**
- Fine Alignment
- Conclusion
- Q & A

Cross-Modal Localization - Coarse Search: Survey

Image-to-LIDAR



Ground RGB (Query) – Aerial LIDAR (Reference)

B. Matei et al., “Image to LIDAR Matching for Geotagging in Urban Environments”, WACV 2013.

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

N. Mithun et al. “RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

.

Image-to-Map

Ground RGB (Query) – OpenStreetMap (Reference)

Castaldo, Francesco, et al. "Semantic cross-view matching." *ICCVW*. 2015.
Panphattarasap et al. “Automated Map Reading: Image Based Localisation in 2-D Maps Using Binary Semantic Descriptors”, IROS 2018.
Samano et al. "You Are Here: Geolocation by Embedding Maps and Images." *ECCV*, 2020.

SRI International

Ground RGB (Query) – Aerial LIDAR (Reference)

- Traditional methods rely on hand-crafted features.
  1. **Very limited prior works** on matching ground RGB images to aerial geo-referenced LIDAR depth data for Cross-Modal Visual Localization [1] [2].
  2. **Limited to urban settings:** (a) performance depends on the **availability of buildings** in the image [1] [2], (b) [1] requires **manually annotated building outlines** of query.
  3. **Evaluated on a very few queries** (14 queries [1] and 50 queries [2]).

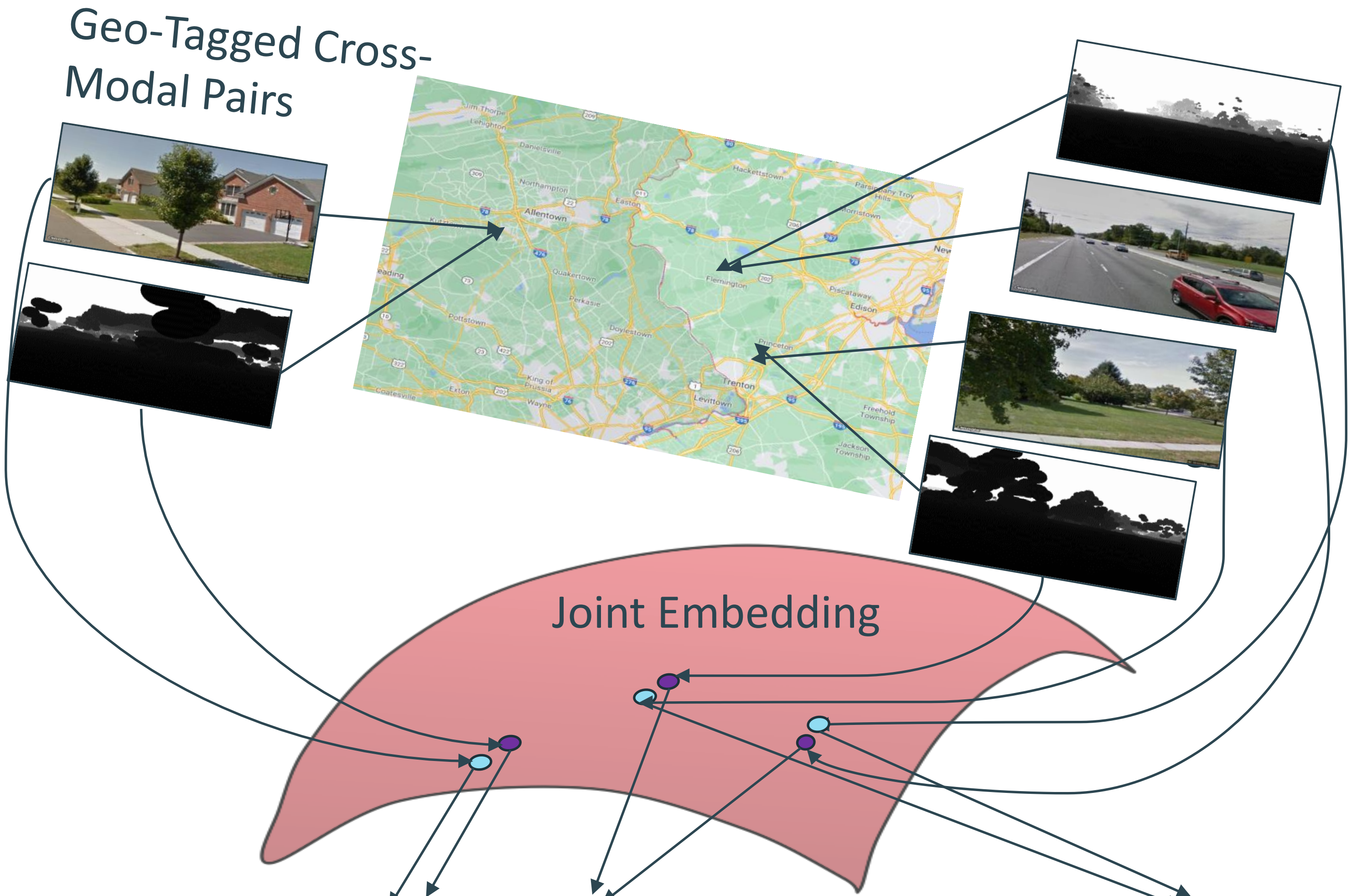
- SRI presented the first deep learning-based approach [3] that utilizes multimodal deep convolutional neural networks (CNN) to learn joint representations for ground-level RGB images and aerial LIDAR depth images.

[1] B. Matei et al., “Image to LIDAR Matching for Geotagging in Urban Environments”, WACV 2013.

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

[3] N. Mithun, K. Sikka, H. Chiu, S. Samarasekera, R. Kumar, “RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020 (Best Paper Finalist).

RGB2LIDAR: Training a Joint 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 Visual Coarse Search

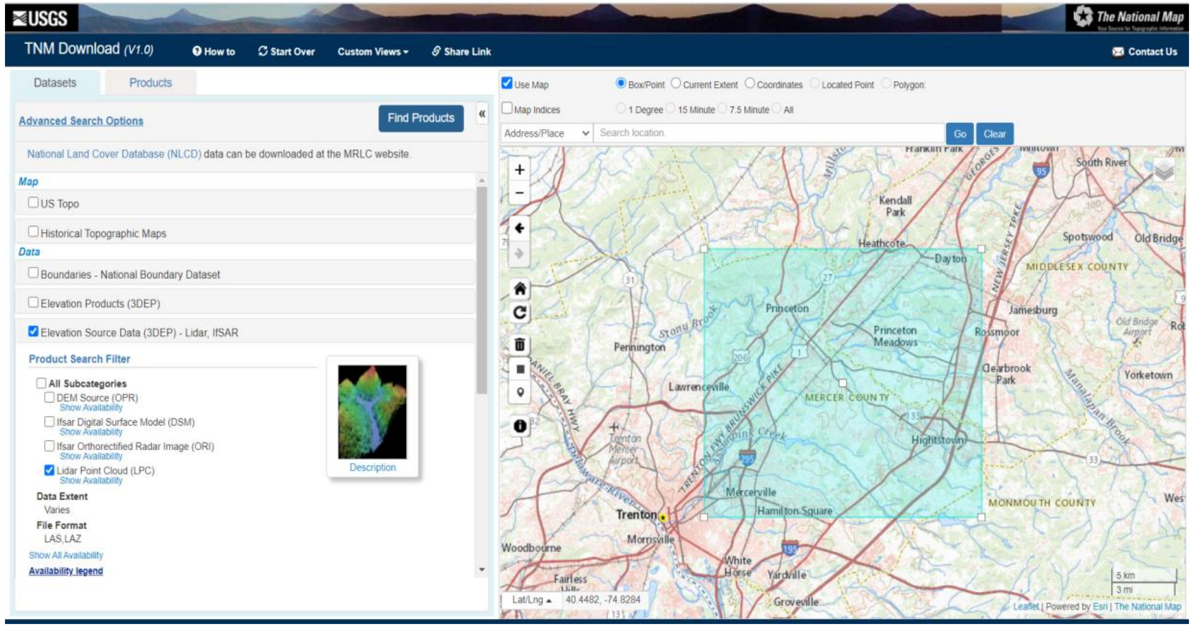
Joint RGB-LIDAR Embedding

- Project 2.5D LIDAR depth images from 3D LIDAR point cloud for different positions and orientations.
- Cross-modal pairs closer in the geo-space should be closer in the embedding space

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

Ground RGB to Aerial LIDAR (GRAL) Dataset

- **Dataset with Location-Coupled Pairs**:
  - **Street-View** Images for different locations (Lat, Long) from Google using Street-View API.
  - **Lidar Depth Images** are collected for the same locations rendering aerial LIDAR 3D point clouds of same area (collected from USGS nationalmap website).



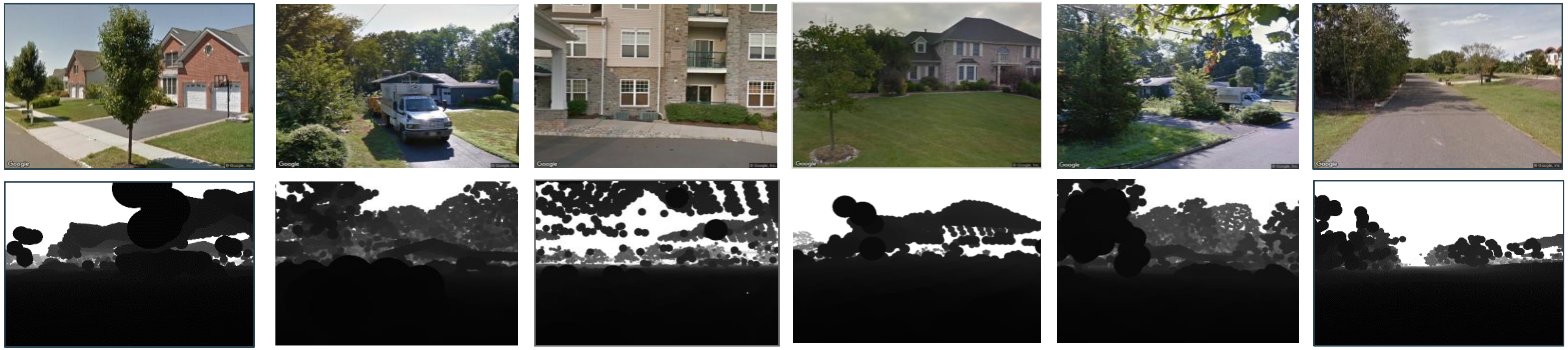
❑ Dataset Available Online at https://github.com/niluthpol/RGB2LIDAR

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

GRAL Dataset

RGB

LIDAR DEPTH



Sample Pairs from the GRAL Dataset

- About ***550K cross-modal pairs*** collected from $143\ \mathrm{km}^2$ area in NJ, USA.
- Weak Alignment between Pairs due to automatic collection (e.g., missing ground pixels in rendered depth images, alignment issue)

Missing Pixels Due to Aerial Collection

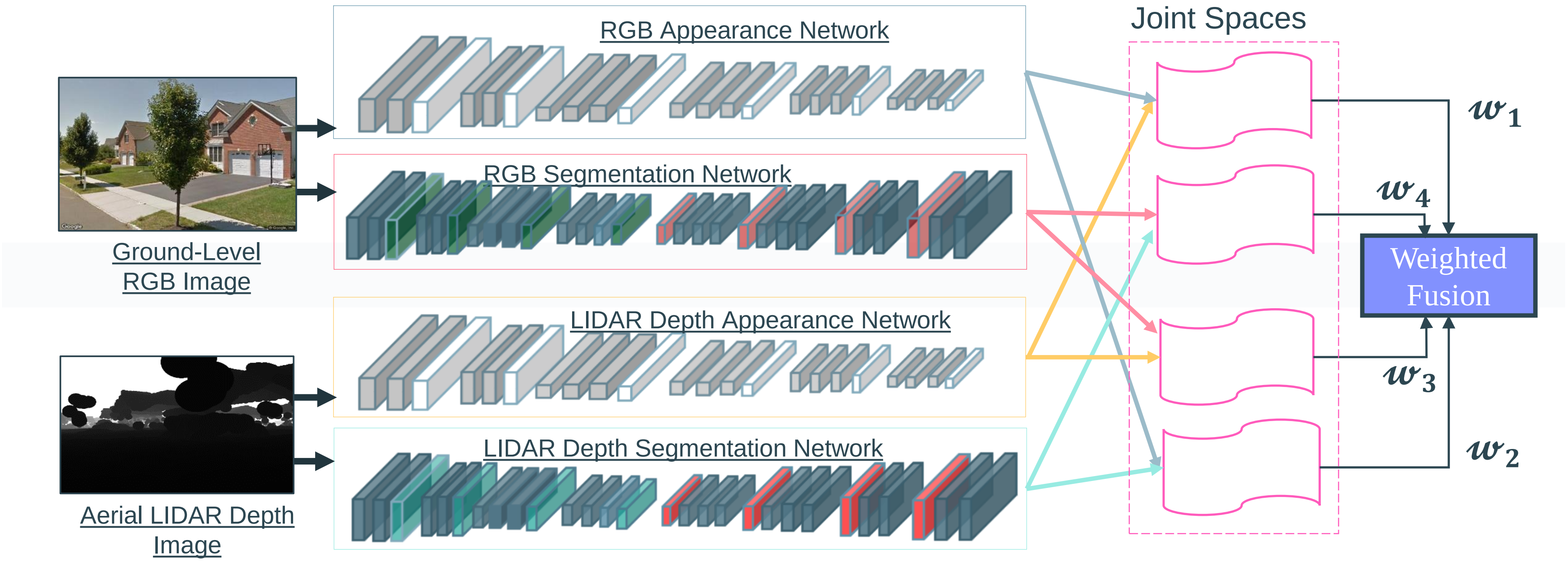
Horizontal and Vertical Alignment

Scene Change over Time

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

Fusion of Appearance & Semantic

- **Both Appearance and Semantic Cue for Retrieval**
  - Matching across modalities exhibits large disparities in appearance characteristics. Higher-level scene information is generally better preserved across inputs, from different visual sensors, capturing the same scene.



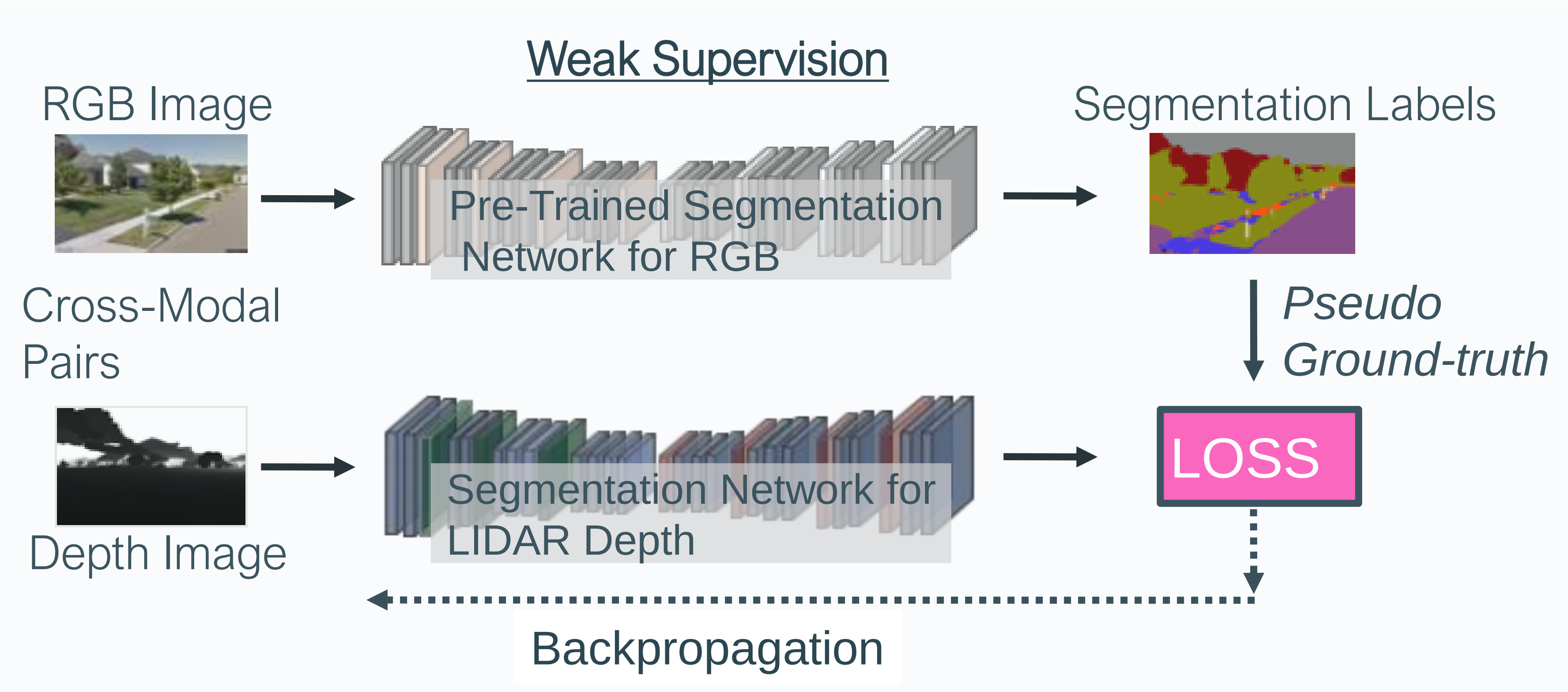
- **Mixture-of-Expert Model for Retrieval**
  - A weighted fusion of joint embedding models trained with different combination of appearance and semantic cues

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

SRI International

Semantic Cues from LIDAR Depth Images

LIDAR Segmentation Network is trained with Weak Cross-Modal Supervision



Approach: Training using the segmentation maps of the paired RGB images as pseudo ground-truth

Quantitative Results

- **Evaluation Metric**
  - R@K (Recall at K) : percentage of queries for which the ground truth results are found within the top-K retrievals.
  - MedR: The median of the ground-truth matches in the ranking.
- **Test Set :** About 50K pairs collected from 14km^2 area
- **Baselines on GRAL Dataset**
  - Baseline hand-crafted or pre-trained CNN models performs slightly better than chance.
  - Result shows difficulty of this cross-modal localization task on the proposed dataset

Method	R@1	R@5	R@10	MedR	MeanR
Chance	0.002	0.010	0.020	24921	24921
GIST	0.002	0.016	0.026	21101	22479
wide-ResNet18	0.014	0.059	0.114	19028	20131
ResNet50	0.007	0.031	0.067	19887	20328
MegaDepth	0.9	3.2	5.1	1735	5628
RGB2LIDAR	27.6	51.1	57.9	5	34.5

- **Proposed RGB2LIDAR Model**
  - Shows promising performance (i.e., R@1 of 27.6 and Median Rank 5).

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

Comparison with Prior Works

- ## Ground RGB to Aerial Lidar based Retrieval

  - Bansal et al.[1] evaluated on **50 queries** and reported **20% accuracy** in $5m$ localization in the **top-1000** ranks in $1Km \times 0.5Km$ area, whereas our method shows **34% in 5$\boldsymbol{m}$** localization in **top-1** testing across **50K** pairs in $143km^2$ area.

  - Matei et al.[2] evaluated their approach on **14 queries in $\boldsymbol{5km^2}$** area and reported **R@1 of 7%,** whereas our method shows **R@1 of 27.6%** based on **50K queries in $\boldsymbol{14km^2}$** area

- ## Ground-Aerial RGB based Retrieval

  - We compare with a prominent cross-view localization model CVM-Net-I [3], by collecting ground panoramas and aerial satellite images for test image locations in GRAL.

  - CVMNet-I model achieves low accuracy (**R@1 =0.7%, R@10 = 5.1%**) in Ground→Aerial-Image localization, whereas our model achieves significantly better (i.e., **R@1 =27.6%, R@10 = 57.9%)** in RGB→LIDAR

[1] M. Bansal et al., "Geometric Urban Geo-Localization", CVPR 2014.
[2] B. Matei et al., "Image to LIDAR Matching for Geotagging in Urban Environments", WACV 2013.
[3] S. Hu et al., "CVMNet: Cross-View Matching Network for Image-Based Ground-to-Aerial GeoLocalization", CVPR 2018

RGB2LIDAR: Analysis

- **Analysis of Proposed Method**
  - Joint Embedding Models trained with all four combinations of appearance and semantic cues from RGB and LIDAR images perform reasonably well.
  - The fusion strategy shows large improvements over single-embedding based baselines
  - Use of LIDAR Depth Semantic feature leads to significant improvements

Method	Evaluation Metric				
	R@1	R@5	R@10	MedR	5m R@1
Chance	0.002	0.01	0.02	24921	0.003
A_R-A_L	20.3	39.0	45.1	19	26.4
S_R-S_L	10.6	26.8	34.8	38	14.1
A_R-S_L	9.5	24.3	32.0	48	12.6
S_R-A_L	18.6	37.2	43.6	22	24.4
A_R-A_L+ S_R-A_L	24.8	45.5	51.8	9	31.5
A_R-A_L+A_R-S_L	22.9	44.7	52.0	9	29.2
A_R-A_L+S_R-A_L+A_R-S_L	26.7	49.5	56.3	6	33.7
A_R-A_L+S_R-A_L+A_R-S_L+S_R-S_L (Proposed)	27.6	51.1	57.9	5	34.5

A_R: Appearance Cues from RGB A_L: Appearance Cues from LIDAR

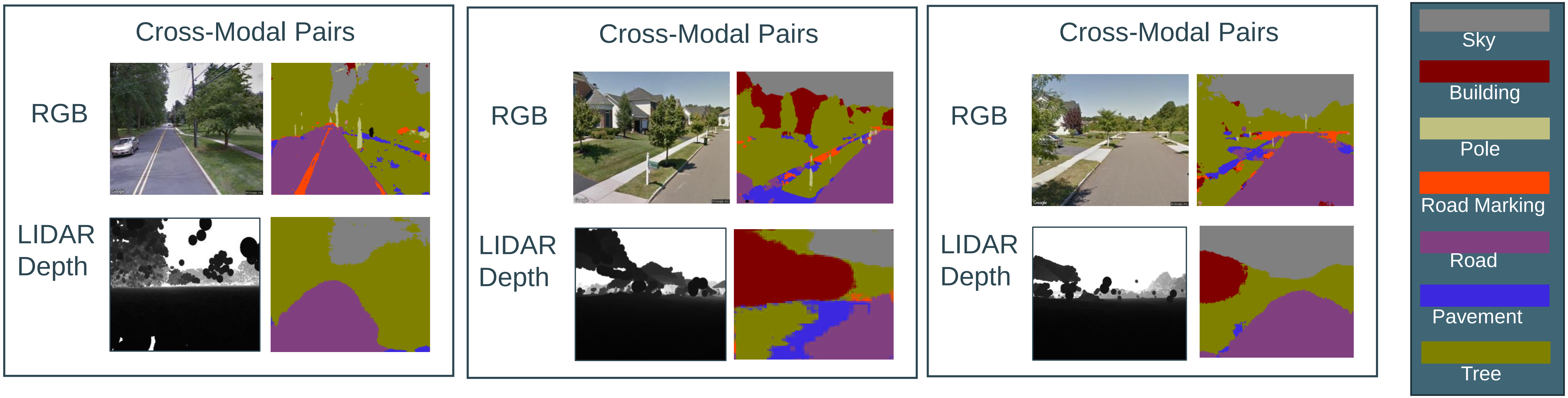
S_R: Semantic Cues from RGB S_L: Semantic Cues from LIDAR

A_R- S_L: Joint Embedding Model trained with Appearance Cues from RGB and Semantic Cues from LIDAR

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

LIDAR Depth Segmentation Results

- LIDAR Depth Segmentation Results are Grounded.
- It support our Intuition of training with Cross-modal Supervision.

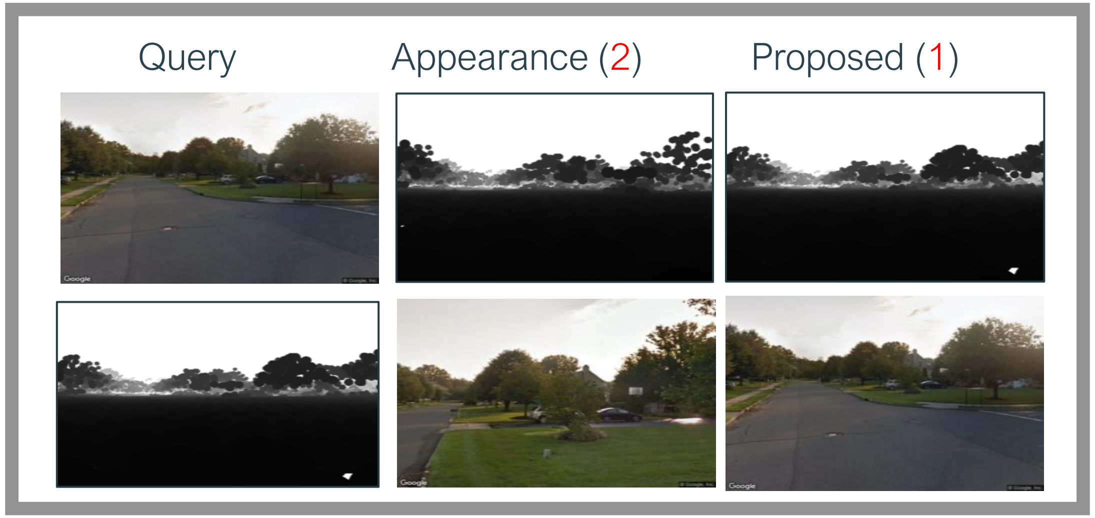


Outputs from Weakly Supervised Segmentation Network

RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

SRI International

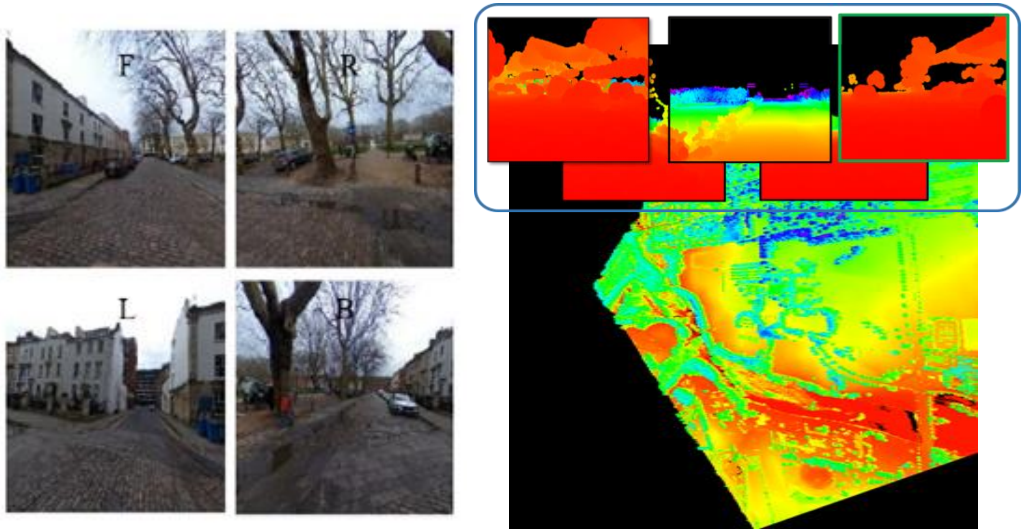
Qualitative Results



RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

Cross-Modal Localization - Coarse Search: Survey

Image-to-LIDAR



Ground RGB (Query) – Aerial LIDAR (Reference)

B. Matei et al., “Image to LIDAR Matching for Geotagging in Urban Environments”, WACV 2013.

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

N. Mithun et al. “RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization, ACM MM, 2020

.

Image-to-Map

Ground RGB (Query) – OpenStreetMap (Reference)

Castaldo, Francesco, et al. "Semantic cross-view matching." *ICCVW*. 2015.

Panphattarasap et al. “Automated Map Reading: Image Based Localisation in 2-D Maps Using Binary Semantic Descriptors”, IROS 2018.

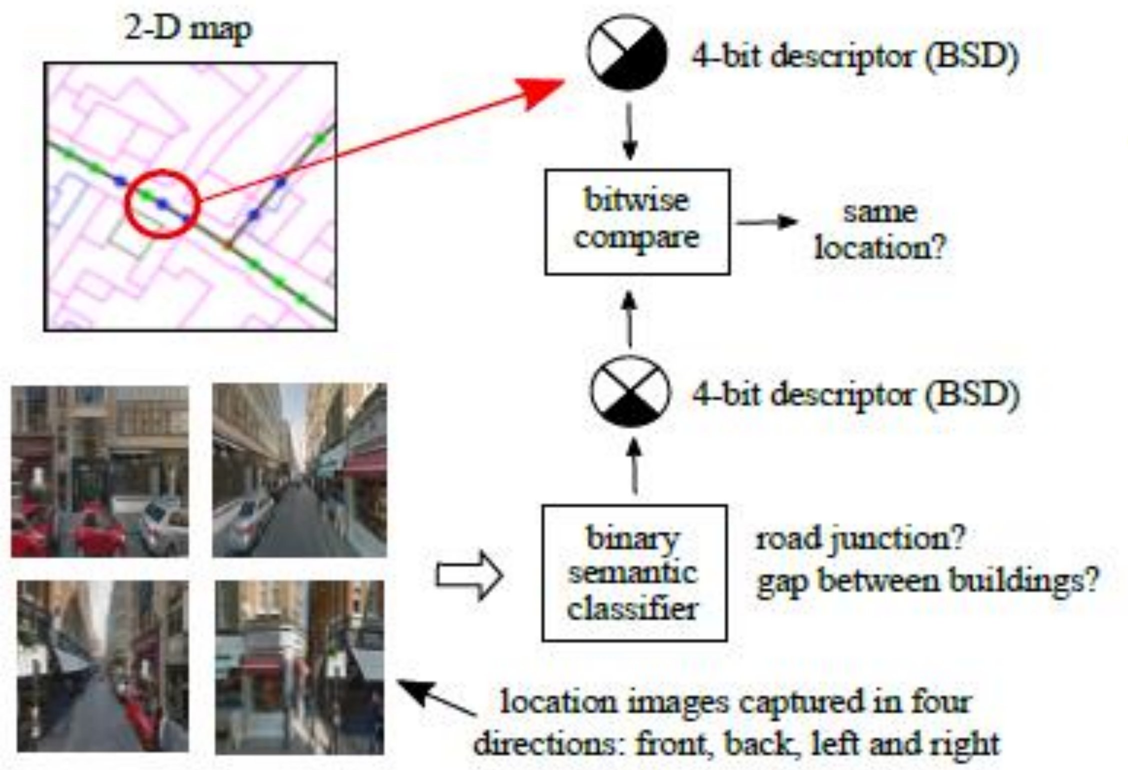
Samano et al. "You Are Here: Geolocation by Embedding Maps and Images." *ECCV*, 2020.

Ground RGB (Query) – OpenStreetMap (Reference)

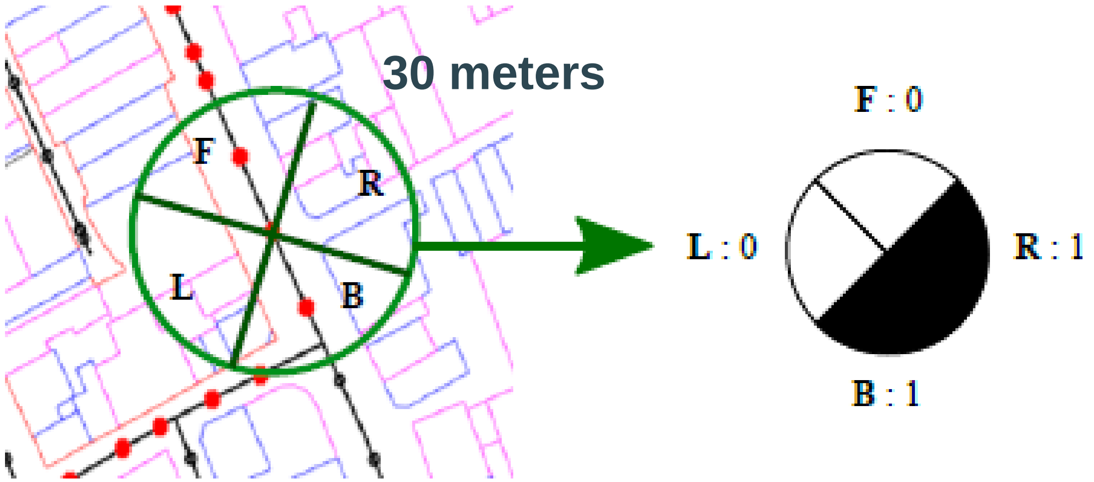
- Related work is limited [1][2][3].
  - Focus on coarse search only - no detailed 3D information in database for fine alignment
- Problem Setting [2][3]:
  - Query Input: Google Street Views (Front, Back, Left, Right) from a panorama image
  - Reference Data: Open Street Map.

[1] Castaldo, Francesco, et al. "Semantic cross-view matching." *ICCVW*. 2015.
[2] Panphattarasap et al. "Automated Map Reading: Image Based Localisation in 2-D Maps Using Binary Semantic Descriptors", IROS 2018.
[3] Samano et al. "You Are Here: Geolocation by Embedding Maps and Images." *ECCV*, 2020.

Automated Map Reading



$$\hat{d}_{ij} = \begin{cases} DETECT_{JUNC}(I_{ij}) & \text{if } j \in \{1,2\} \\ DETECT_{BGAP}(I_{ij}) & \text{if } j \in \{3,4\} \end{cases}$$



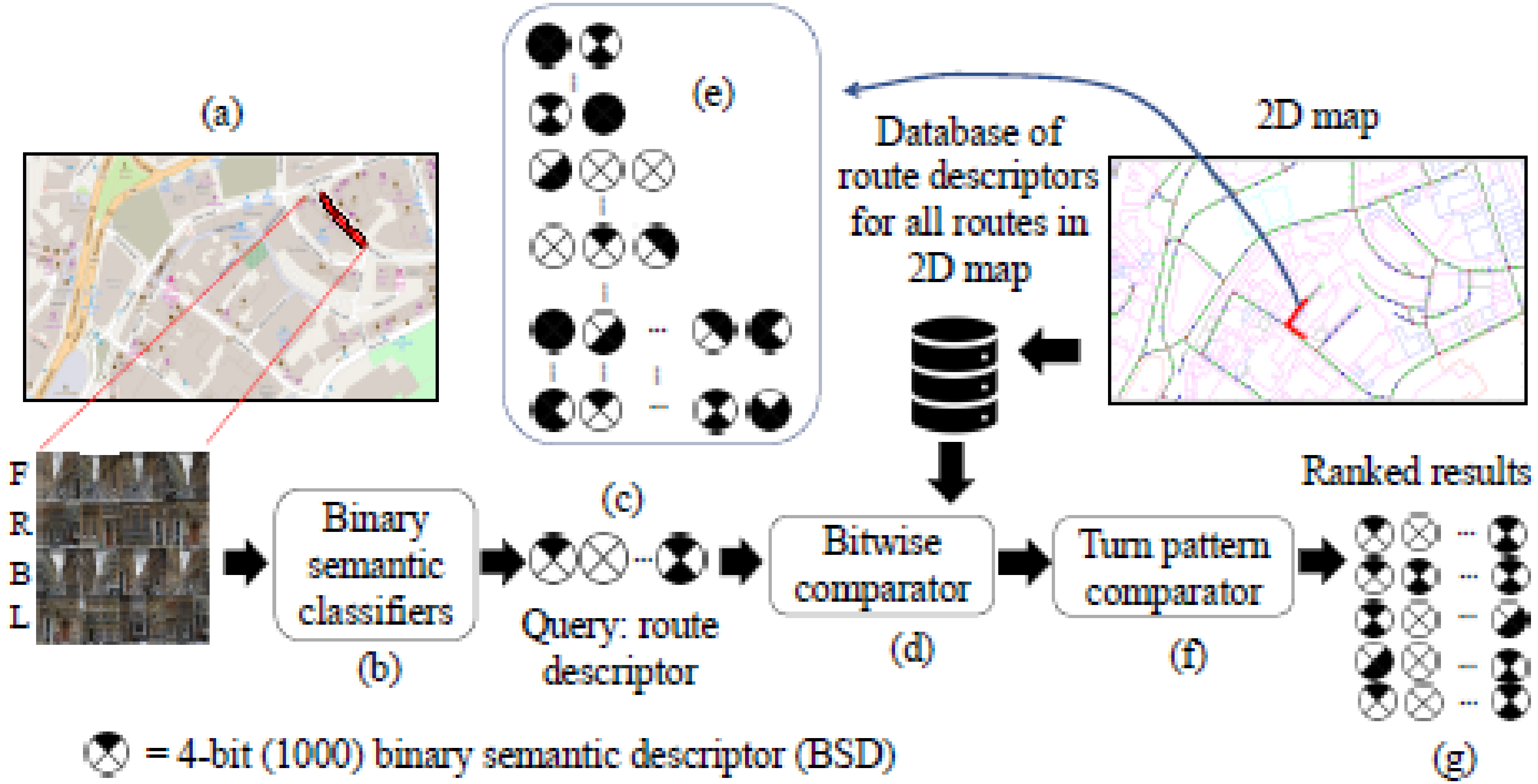
Panphattarasap et al. "Automated Map Reading: Image Based Localisation in 2-D Maps Using Binary Semantic Descriptors", IROS 2018

'0000' :no gaps and no junctions

- Fine-tune an off-the-shelf pre-trained CNN (Places205-AlexNet model) using paired GSV-OSM data.
  - A training set of 440,000 images per classifier taken from 220,000 locations in 23 different cities in the UK.
  - **75% accuracy for two test sets of 8000 images taken from the same 23 cities**

Route Descriptor and Turn Pattern

- Route Descriptor: Connect positions every 10 meters
- Turn Pattern: whether a left and right turn (>60 degree) presents between positions



SRI International

Automatic Map Reading: Image Localisation in 2-D Maps Using Binary Semantic Descriptors", IROS 2018.

Automated Map Reading: Experimental Results

- Using GSV and OSM data for a 2.5 km^2 region of London. The region consisted of 6656 GSV locations.
- 150 test routes, maximum route length (40 locations, 400 meters)

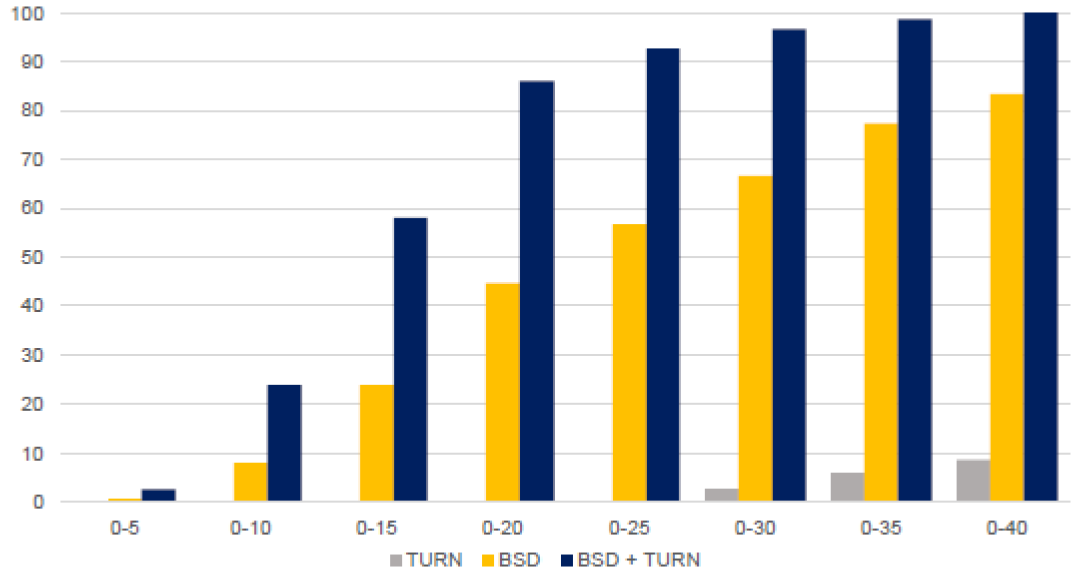


Fig. 7. Accuracy of localisation (% of correctly identified routes) versus route length using turn patterns (grey), route descriptors (yellow), and route descriptors with turn patterns (blue).

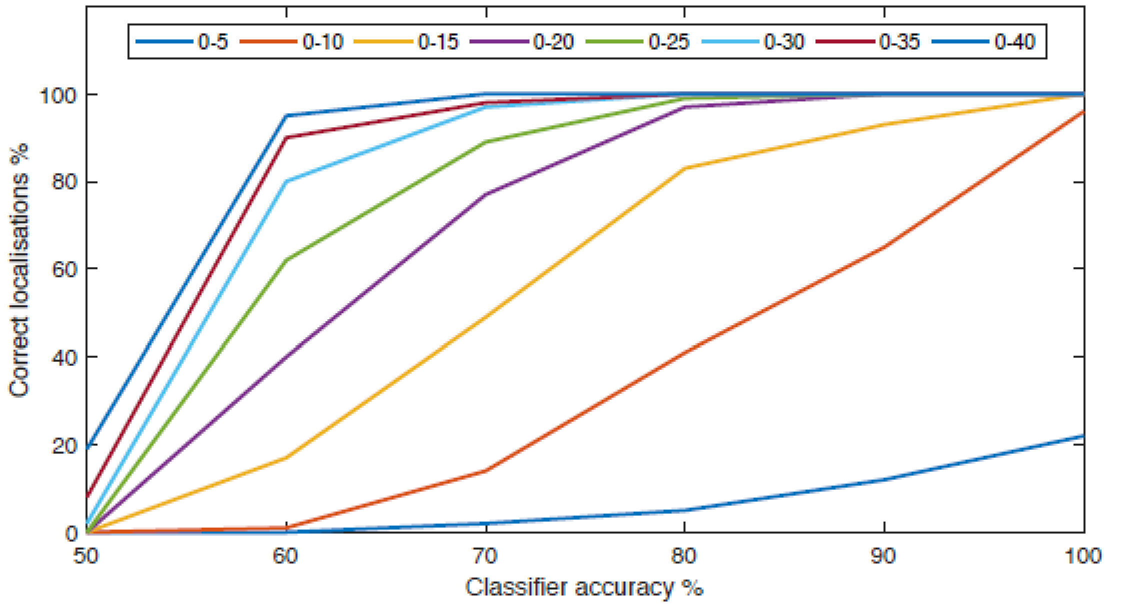


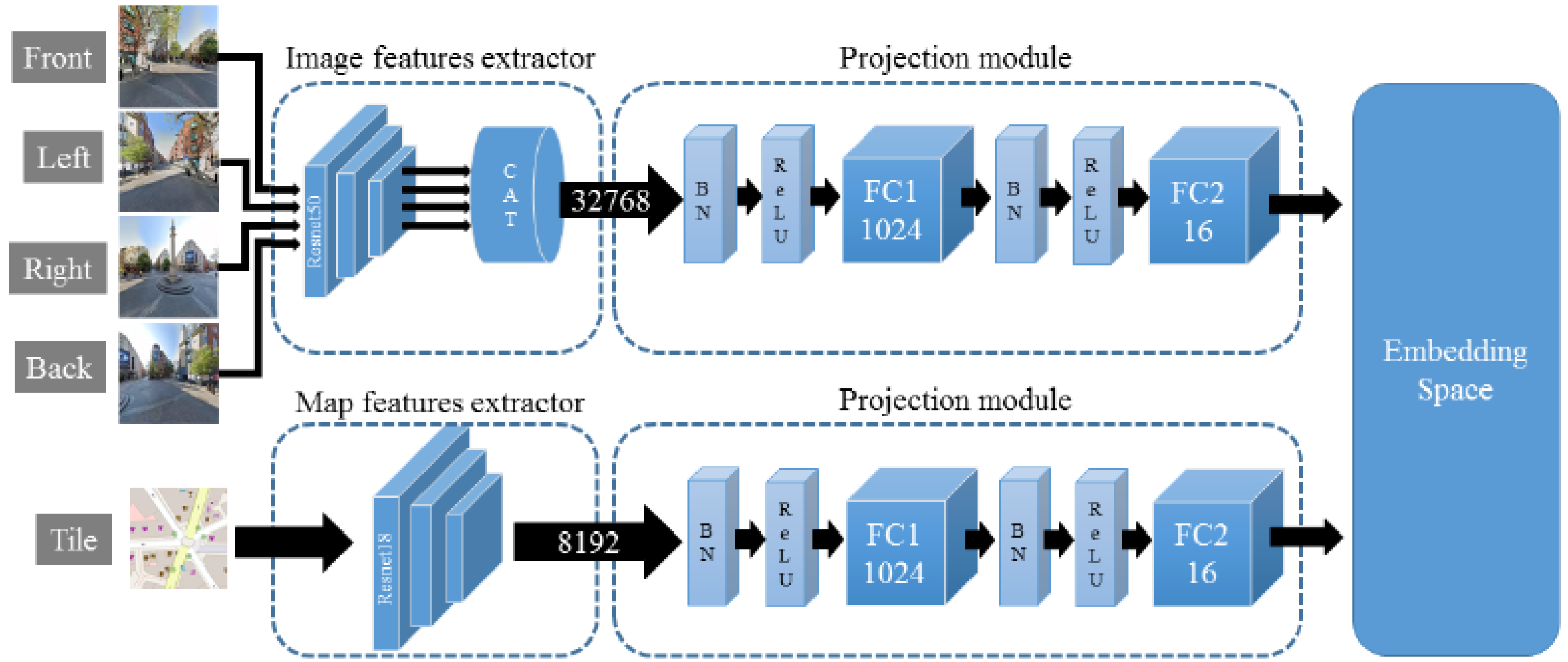
Fig. 8. Accuracy of localisation (% of correctly identified routes) versus classifier accuracy for different ranges of route length.

SRI International

Automatic Map Reading: Image Localisation in 2-D Maps Using Binary Semantic Descriptors", IROS 2018.

Geolocation by Embedding Maps and Images

- Feature Extraction: Image (Resnet 50), 4*4 feature map of 512-d vector, combine 4 inputs (32768)
- Feature Extraction: Map (Resnet18, fewer details), 4*4 feature map of 512-d vector
- Projection module: Two fully connected layers, both preceded by batch normalization and ReLu activation. (16-d)



Samano et al. "You Are Here: Geolocation by Embedding Maps and Images." ECCV, 2020.

Training set consisted of 98,767 panorama images and two tiles (152m*152 m - S1 and 76m*76m - S2) for each location.

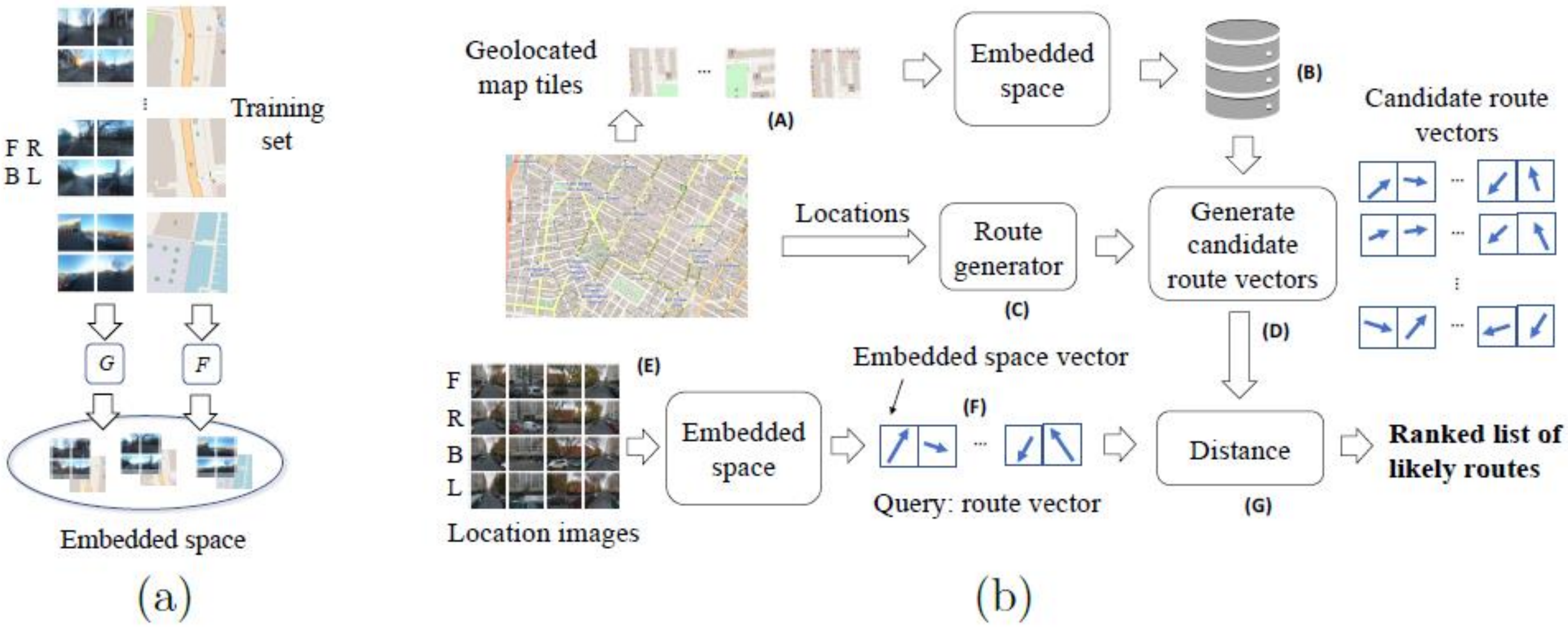
Data Augmentation for Training: Small changes in the scale of the map tiles and the viewing directions when cropping the panoramic images to form triplets (examples of matched and un-matched image/map tile pairs inside every batch).

Triplet Loss function

You Are Here: Geolocation by Embedding Maps and Images. *ECCV*, 2020

You Are Here: System Pipeline

- Use the same route descriptor concept to improve the discrimination.



You Are Here: Geolocation by Embedding Maps and Images. *ECCV*, 2020

You Are Here: Experimental Results

- The StreetLearn data set, which contains 113,767 panoramic images extracted from GSV in the cities of New York (Manhattan) and Pittsburgh.
- Three testing data sets: each with 5,000 panoramas and 10,000 map tiles.

SRI International

You Are Here: Geolocation by Embedding Maps and Images. *ECCV*, 2020

Outline

- Cross-Modal Geo-Localization
- Coarse Search
- **Fine Alignment**
- Conclusion
- Q & A

Cross-Modal Localization – Fine Alignment: Survey

Direct 2D-3D Registration

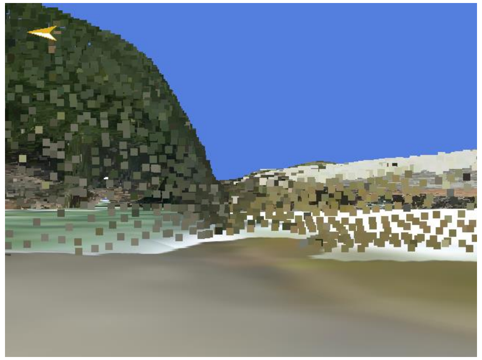
2D Image

3D Point Cloud

Nadeem et al., Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Cloud," ICONIP 2019

Registration via 2.5D Rendering

2D Image



2.5D Rendered View

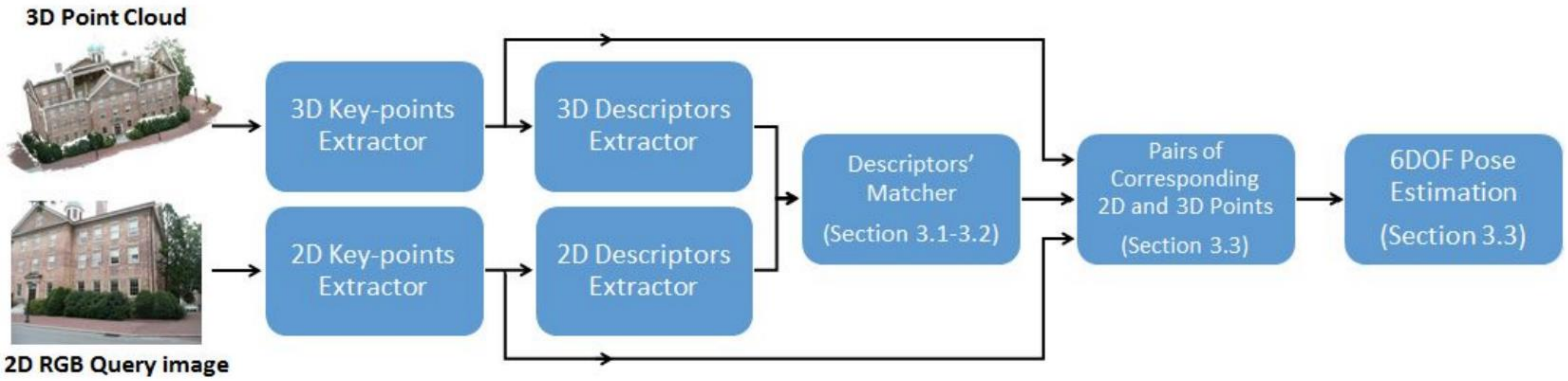
Chiu et al., Augmented Reality Driving Using Semantic Geo-Registration," IEEE Virtual Reality 2018.

Cattaneo et al., CMRNet: Camera to LIDAR-MAP Registration, ITSC, 2019.

Cattaneo et al., CMRNet++: Map and Camera Agnostic Monocular Visual Localization in LiDAR Maps, ICRA 2020.

Direct 2D-3D Registration

- Typical direct 2D-3D registration methods assume the 3D point cloud constructed using structure-from-motion techniques (from multiple camera images).
  - 3D representation has 2D information (such as 2D appearance) from images
- For 3D point cloud from different modalities (such as LIDAR), [1] trains a random forest classifier to match 2D and 3D descriptors.



[1] Nadeem et al., Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Cloud," ICONIP 2019

Cross-Modal Localization – Fine Alignment: Survey

Direct 2D-3D Registration

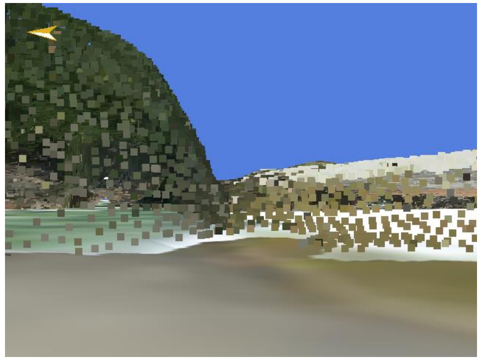
2D Image

3D Point Cloud

Nadeem et al., Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Cloud," ICONIP 2019

Registration via 2.5D Rendering

2D Image



2.5D Rendered View

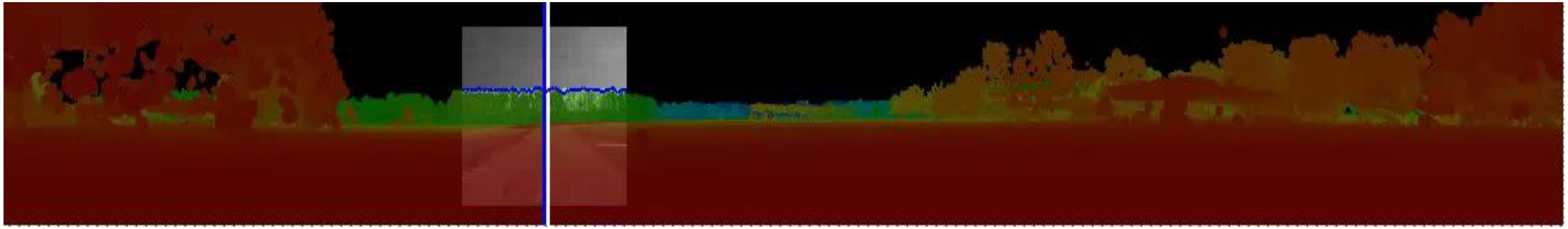
Chiu et al., Augmented Reality Driving Using Semantic Geo-Registration," IEEE Virtual Reality 2018.

Cattaneo et al., CMRNet: Camera to LIDAR-MAP Registration, ITSC, 2019.

Cattaneo et al., CMRNet++: Map and Camera Agnostic Monocular Visual Localization in LiDAR Maps, ICRA 2020.

Registration via 2.5D Rendering

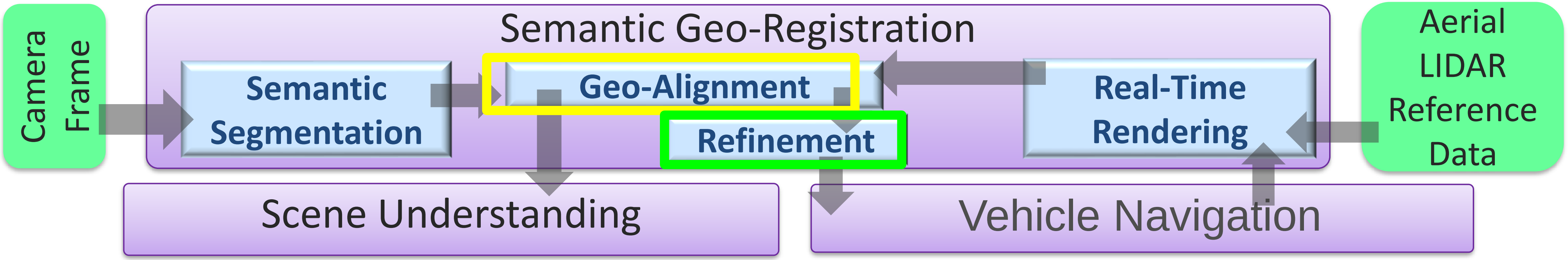
- Render 2.5D view from 3D RGBD point cloud using initial pose
  - Leverage available cross-time registration techniques for 3D pose refinement
  - Refer to cross-time fine-alignment session in this tutorial
- If no appearance information in 3D point clouds …
  - Image-Depth registration becomes difficult
  - There are works using traditional semantic features (such as skyline, building outline) for matching and registration.

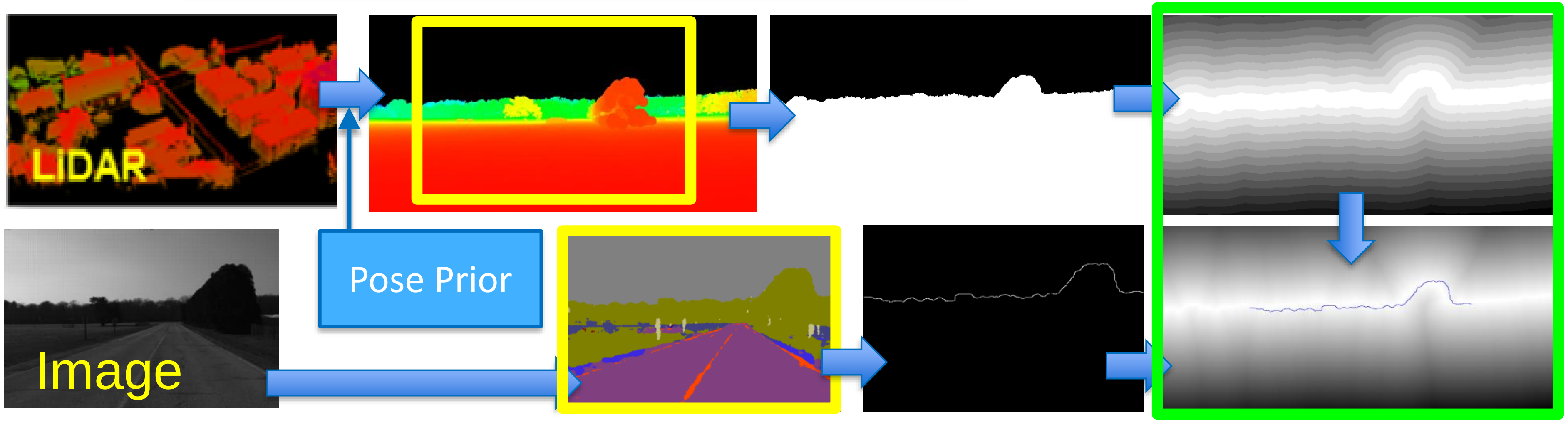


Chiu et al. "Augmented Reality Driving Using Semantic Geo-Registration." IEEE Virtual Reality, 2018.

SRI International

Semantic Geo-Registration





Refinement is only successful when sufficient portions of skyline (75%) are observed.

$$\arg\min_{k} \sum_{n=1}^{N} T(i_n, j_n) D(i_n + k, j_n),$$

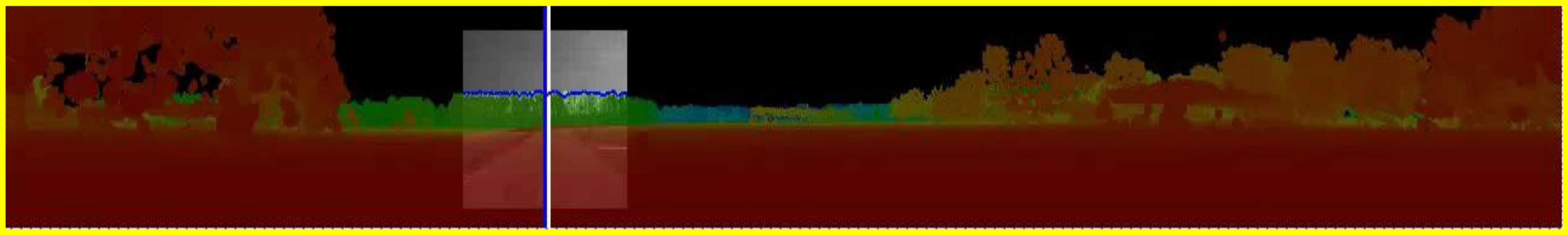
- **Modified Chamfer Matching method:** Find best alignment of template T over D, by summing up distance transform values for all N skyline pixels on T

SRI International

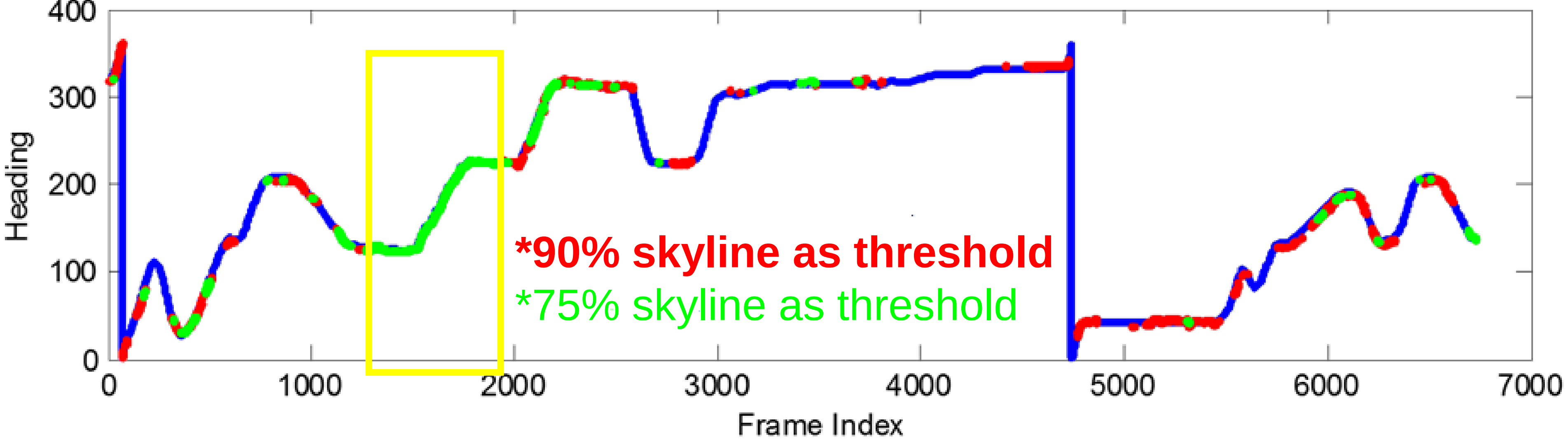
Augmented Reality Driving Using Semantic Geo-Registration. *IEEE VR*, 2018

Heading Refinement

- Rural area: 20~50 mph (max 60 mph).
- Urban city: Slow speed (10~20 mph) due to traffic.
- Skyline refinement is shown on a 360 degree 2.5D rendered depth map.



Training Field: 1.29 degree error, > 50% successful rate.

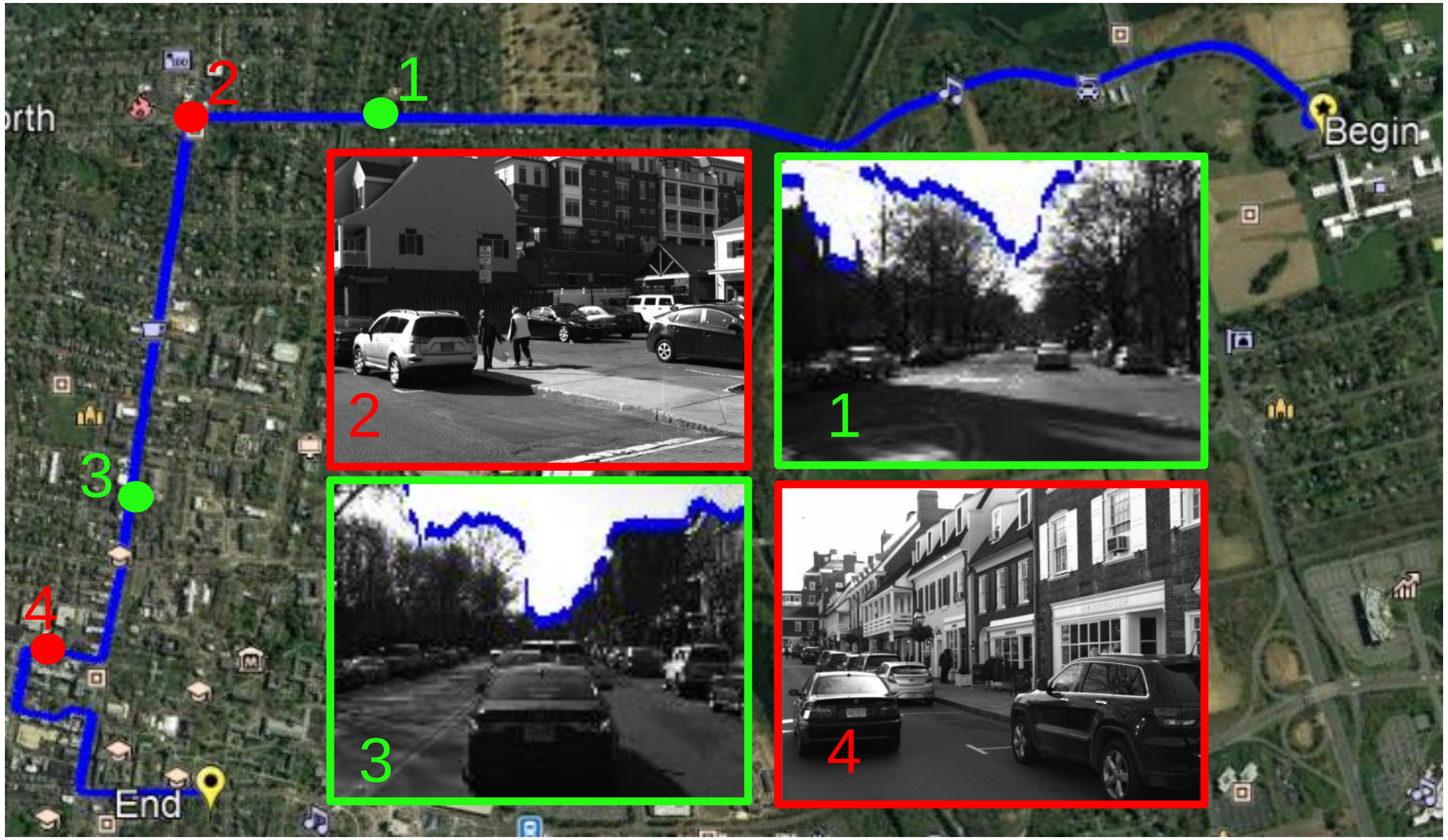


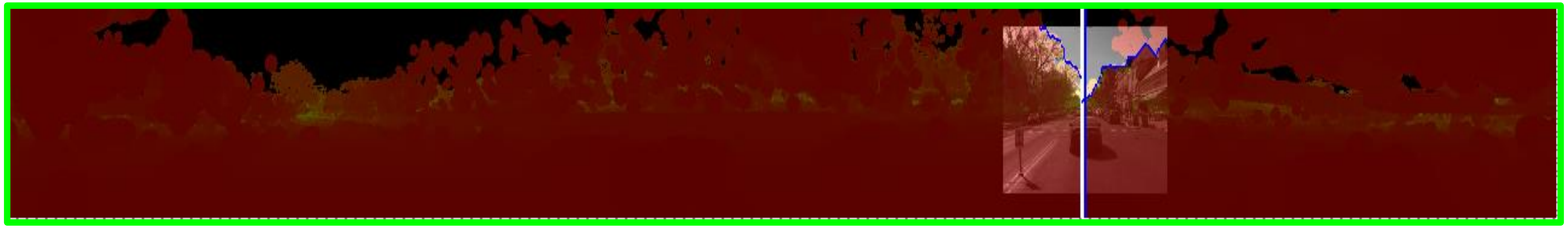
Augmented Reality Driving Using Semantic Geo-Registration. *IEEE VR*, 2018

Heading Refinement: Failure Cases

Urban city:

- Successful rate decreases to 36.97%.
- The median heading error is 0.985 degree.

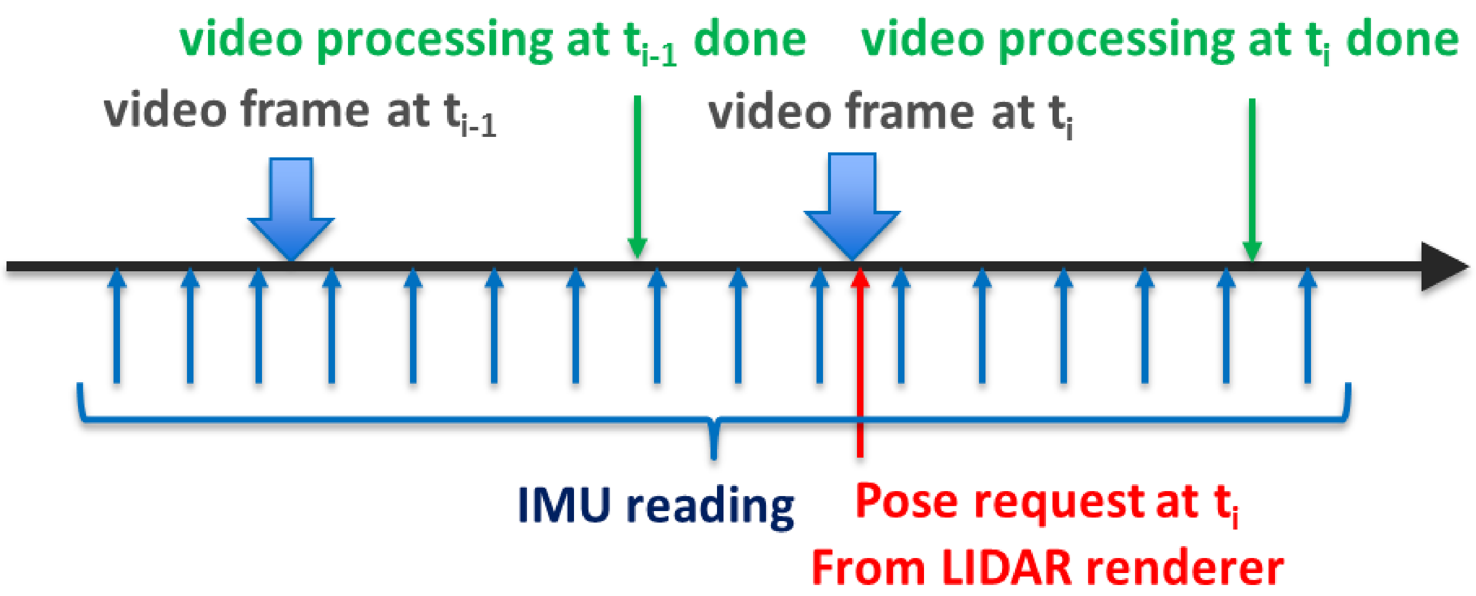




SRI International

Augmented Reality Driving Using Semantic Geo-Registration. *IEEE VR*, 2018

Semantic Geo-Registration for Navigation

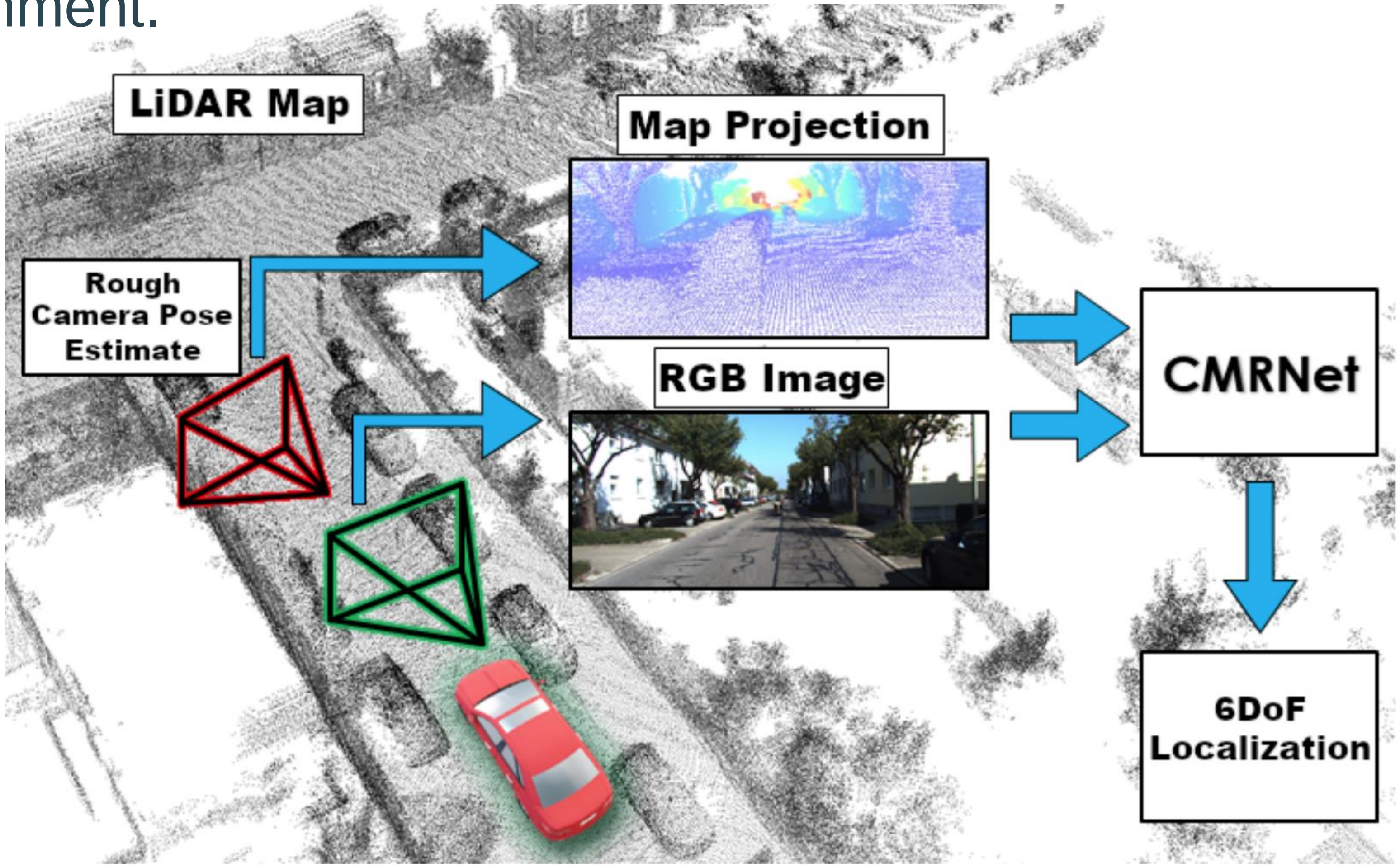


- IMU Pre-Integrated Mechanism
- Visual Odometry and GPS
- Global Heading Update: We propagate opportunistic heading corrections through IMU dynamics over time to improve overall accuracy

Augmented Reality Driving Using Semantic Geo-Registration. *IEEE VR*, 2018

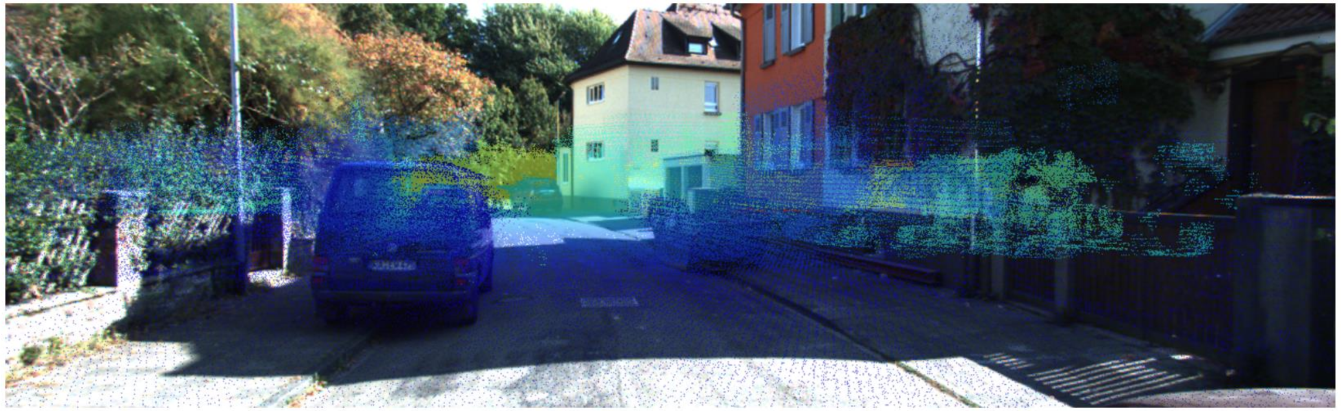
CMRNet: Camera to LiDAR-Map Registration

- The first end-to-end deep learning pipeline for image-depth registration to 3D pose fine-alignment.



Cattaneo et al., CMRNet: Camera to LIDAR-MAP Registration, ITSC, 2019.

CMRNet: LiDAR-image Generation



(a) Without Occlusion Filter

(b) With Occlusion Filter

- Render LiDAR-image (depth image) using an initial pose (from coarse search).

$$H_{init}$$

- Project 3D points into image plane:

$$p^i = K \cdot H_{init} \cdot P^i$$

- Apply occlusion filter

CMRNet. *ITSC, 2019*

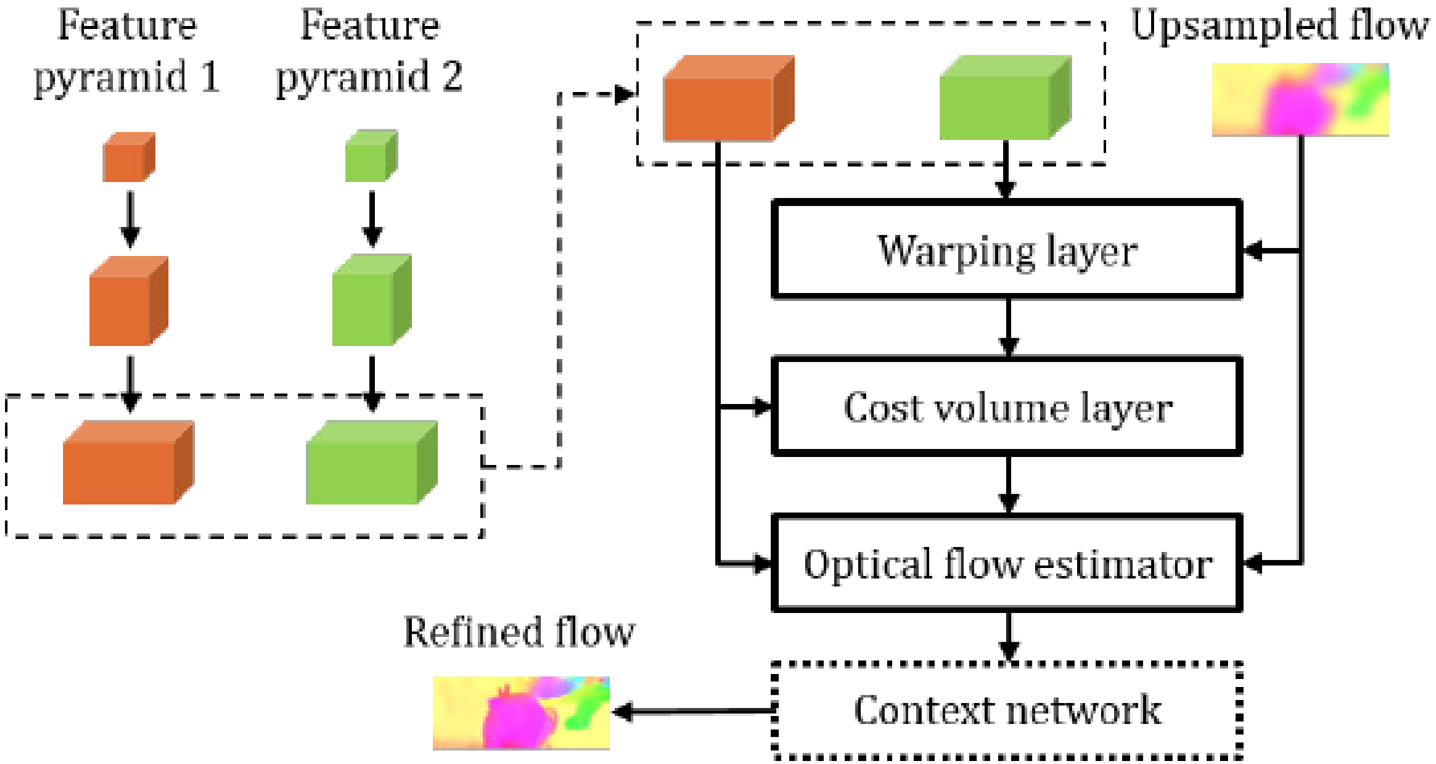
CMRNet: Network Architecture from PWC-Net

Network architecture:

- Two branches of encoder for RGB and depth images
- Decouple the feature pyramid extractors by removing the weights sharing.
- Remove the up-sampling layers and attach the fully connected layers after the first cost volume layer.
- Two branches for rotations and rotations after flow estimation

Loss:

$$\mathcal{L}(\mathcal{I}, \mathcal{D}) = \mathcal{L}_t(\mathcal{I}, \mathcal{D}) + \mathcal{L}_q(\mathcal{I}, \mathcal{D})$$



PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume, CVPR, 2018.

CMRNet. *ITSC, 2019*

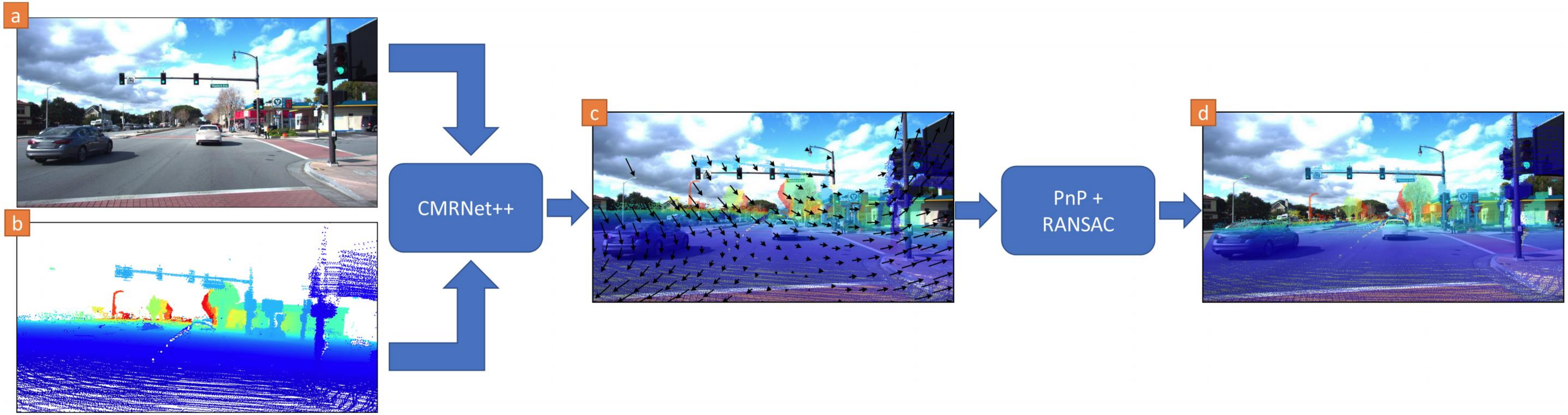
CMRNet: Results

- Achieve 0.27meter and 1.07degree accuracy, starting from initial pose within 3.5meter and 17degree error range.

CMRNet. *ITSC, 2019*

CMRNet++: Map and Camera Agnostic Monocular Visual Localization in LiDAR Maps

- CMRNet++ uses PWC-Net as the backbone network for flow prediction.
- Train CMRNet++ to predict pixel displacement
- Run RANSAC based on point matches predicted by the CMRNet++.



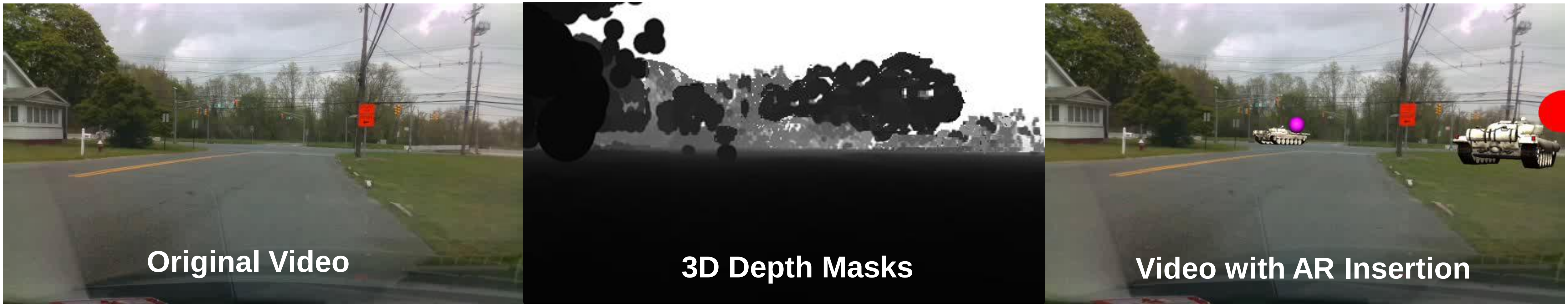
CMRNet++, ICRA, 2020

Outline

- Cross-Modal Geo-Localization
- Coarse Search
- Fine Alignment
- **Conclusion**
- **Q & A**

Image-Based Cross-Model Geo-Localization

- Challenging due to large difference in appearance across modalities
- Huge potential and broad impact to many applications
- Limited works in utilizing deep learning for this problem
- ***Great research direction and topics for exploration!***



*This example is from SRI ONR WAR3D project

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