Vision-based Metric Cross-view Geolocalization

CVPR 2023: A Comprehensive Tour and Recent Advancements toward Real-world Visual Geo-Localization

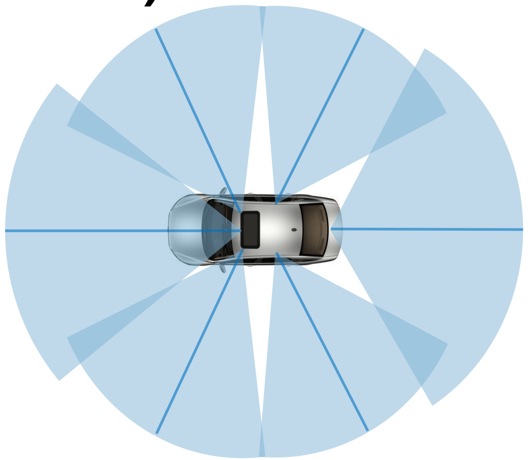
Florian Fervers

florian.fevers@iosb.fraunhofer.de fferflo.github.io

Images: CVUSA [1]

Problem: Cross-view Geolocalization (CVGL)

Input:



1. Ground: Visual, lidar, radar sensors
2. Aerial: Visual, semantic, infrared, elevation orthomaps

Map data: Bing Maps 2023, © Vexcel Imaging

Output:

Georegistered location (+orientation)

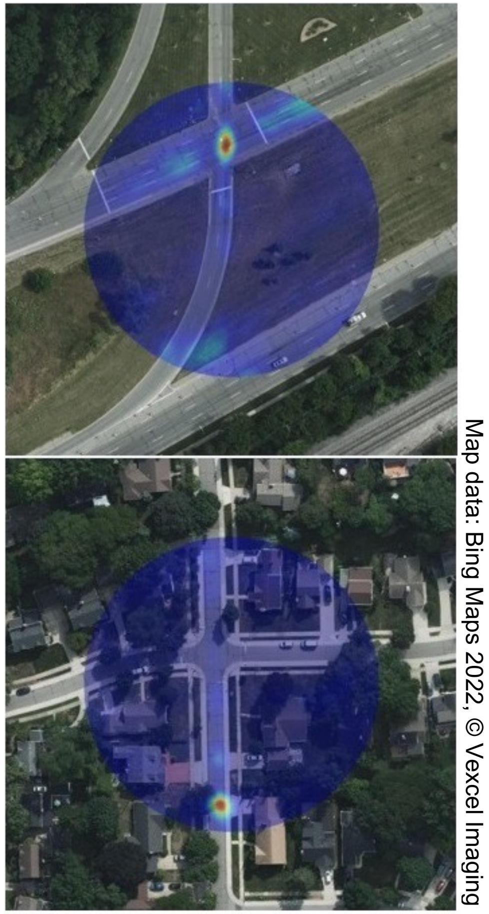
Problem: Cross-view Geolocalization (CVGL)

Two categories of approaches:

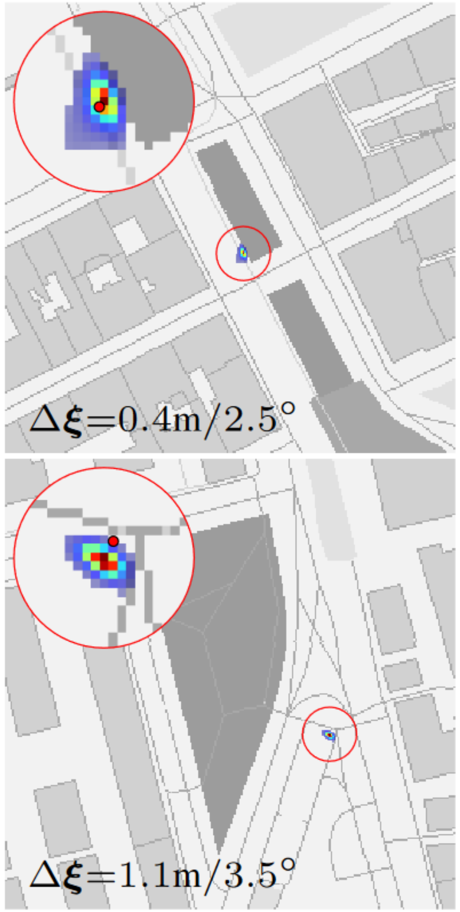
	Large-Area CVGL	Metric CVGL
Search region	Large (e.g. city-scale)	Small (< ~100m)
Approach	Image Retrieval	Pose estimation
Prediction	Target image patch (~10-100m) Probabilistic	Metric pose (Non-)Probabilistic
Metrics	Recall	Recall, mean position error
Datasets	CVUSA [1], CVACT [2], VIGOR [3], …	???

Metric Cross-view Geolocalization

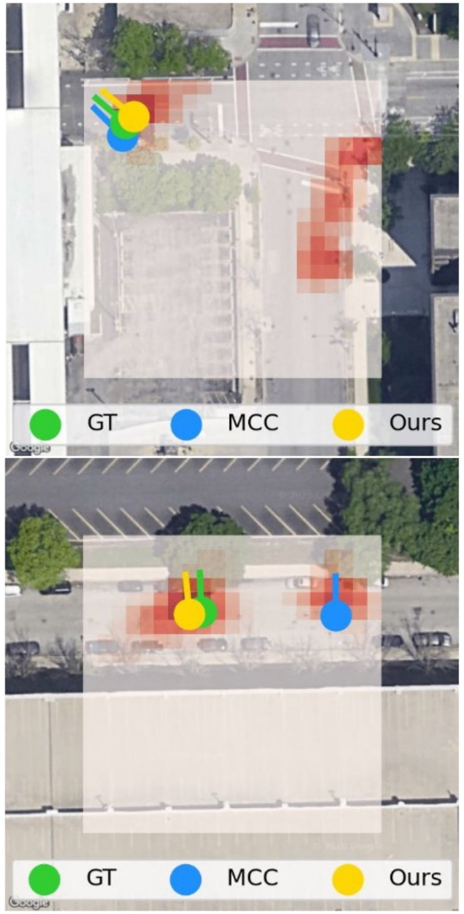
Example predictions from CVPR2023 papers:



Ours [4]



OrienterNet [5]



SliceMatch [6]

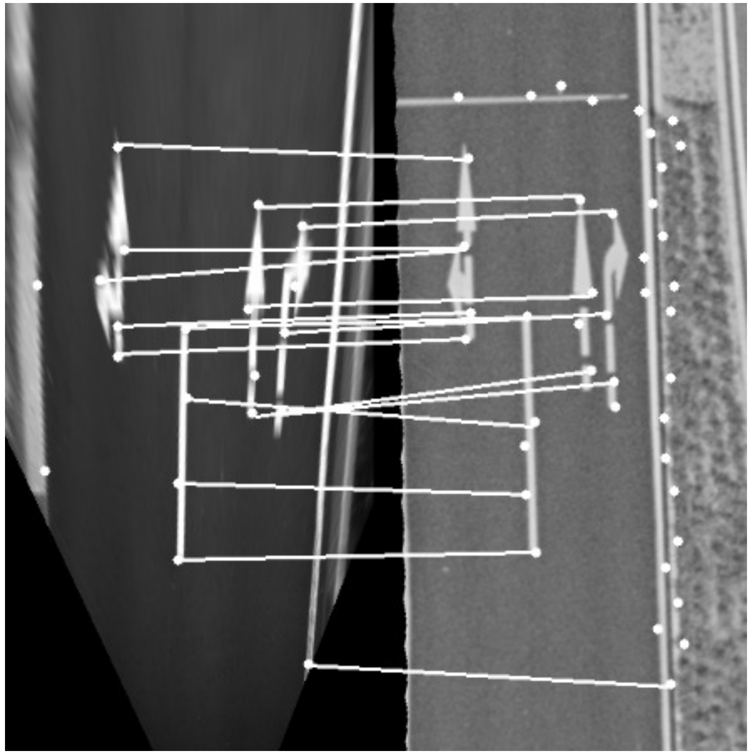
Metric Cross-view Geolocalization

Categories of approaches based on extracted features:

Features	Properties
Local feature descriptors (e.g. SURF [7]) Raw data (e.g. NMI [8])	– Invariance ↔ Discriminance – Unmatched surface areas – Transformation between PV and BEV
Semantic: Buildings [9,10], roads + trajectory [11,12], lane markings [13,14], vertical structures, …	– Invariance ↔ Discriminance – Requires presence of semantic classes – Transformation between PV and BEV
End-to-end learned [4,5,6,17,18]	+ Invariance ↔ Discriminance + Transformation between PV and BEV can be learned – Data and ground-truth collection

Example: Local feature descriptors

1. Project to BEV via homography
2. Extract & match SURF features



Projected ground image   Aerial image

From: *Vehicle ego-localization by matching in-vehicle camera images to an aerial image* (Noda et al., 2011) [7]

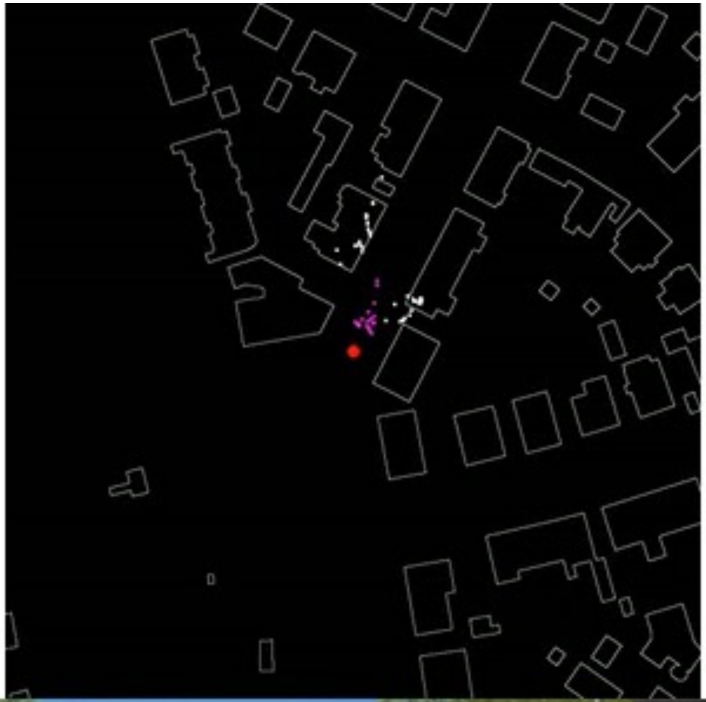
Metric Cross-view Geolocalization

Categories of approaches based on extracted features:

Features	Properties
Local feature descriptors (e.g. SURF [7]) Raw data (e.g. NMI [8])	– Invariance ↔ Discriminance – Unmatched surface areas – Transformation between PV and BEV
Semantic: Buildings [9,10], roads + trajectory [11,12], lane markings [13,14], vertical structures, …	– Invariance ↔ Discriminance – Requires presence of semantic classes – Transformation between PV and BEV
End-to-end learned [4,5,6,17,18]	+ Invariance ↔ Discriminance + Transformation between PV and BEV can be learned – Data and ground-truth collection

Example: Prototype

1. Visual SLAM (ORB-SLAM)
2. Semantic segmentation
3. Iterative closest points



Vehicle Data: KITTI Map data: OpenStreetMaps

Metric Cross-view Geolocalization

Categories of approaches based on extracted features:

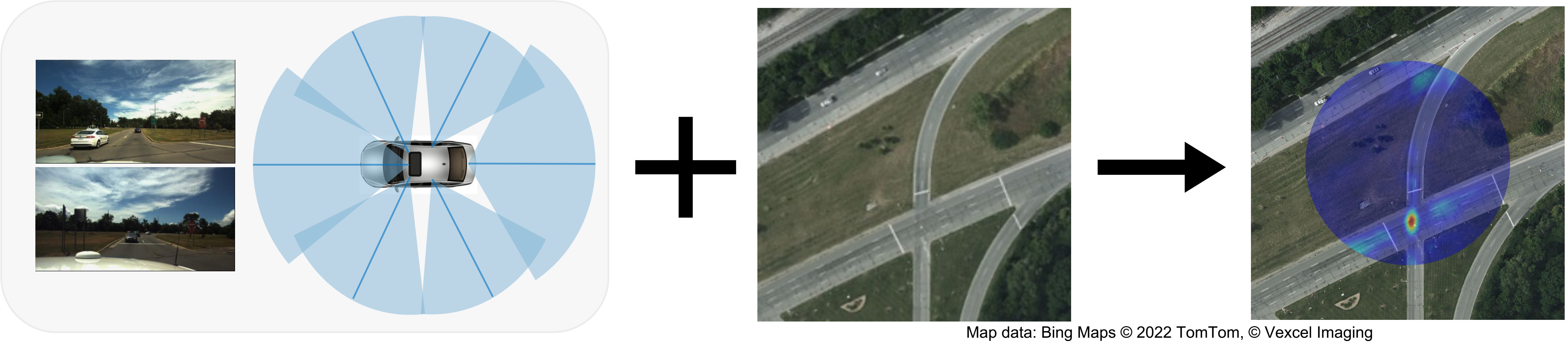
Features	Properties
Local feature descriptors (e.g. SURF [7]) Raw data (e.g. NMI [8])	– Invariance ↔ Discriminance – Unmatched surface areas – Transformation between PV and BEV
Semantic: Buildings [9,10], roads + trajectory [11,12], lane markings [13,14], vertical structures, …	– Invariance ↔ Discriminance – Requires presence of semantic classes – Transformation between PV and BEV
End-to-end learned [4,5,6,17,18]	+ Invariance ↔ Discriminance + Transformation between PV and BEV can be learned – Data and ground-truth collection

End-to-end Metric CVGL

1. With range-scanners
   - First: *Rsl-net: Localising in satellite images from a radar on the ground*
     Tang et al. RA-L 2020
   - Ours: *Continuous self-localization on aerial images using visual and lidar sensors*
     Fervers et al. IROS 2022

2. Vision-only (without range-scanners)
   - Related: Image retrieval methods [19][20], regression [3]
   - First: *Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image*
     Shi et al. CVPR 2022
   - Ours: *Uncertainty-aware Vision-based Metric Cross-view Geolocalization*
     Fervers et al. CVPR 2023

Uncertainty-aware Vision-based Metric Cross-view Geolocalization, *Fervers et al., CVPR 2023 [4]*



Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Main Contributions:

1. Propose end-to-end trainable model for vision-based metric CVGL
2. State-of-the-art performance even in zero-shot setting
3. Improved ground-truth for multiple datasets

Code and ground-truth available at https://fferflo.github.io/projects/vismetcvgl23

Model Summary

(a) Feature extraction

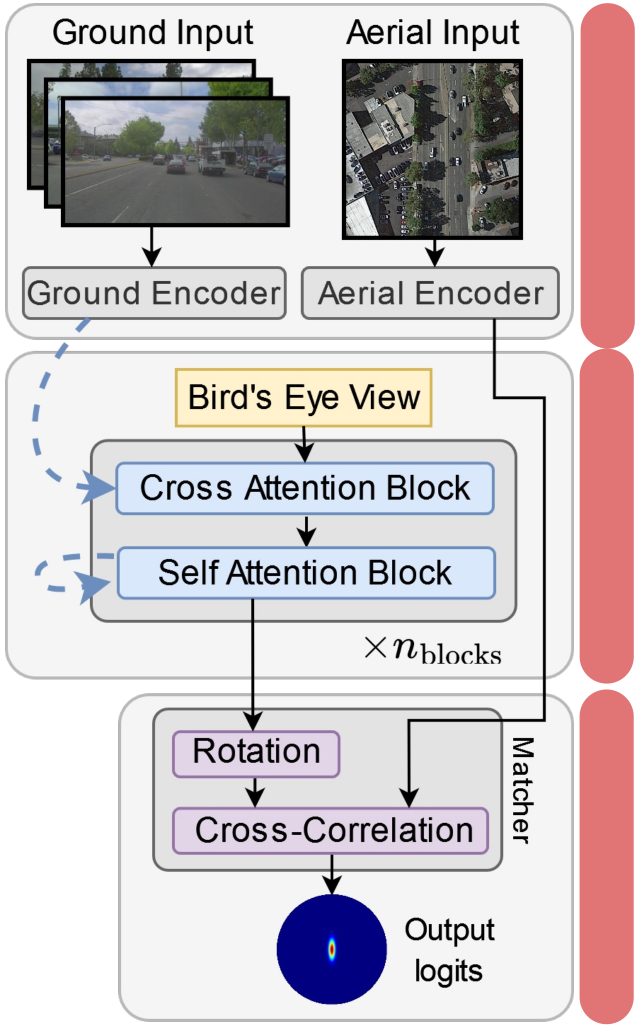
- ConvNeXt [1] + simple decoder
- Shared weights for ground images

(b) Perspective View to Bird's Eye View (PV2BEV)

- Cross-attention: BEV point pillars projected onto PVs (with deformable offsets)
- Self-attention: SegFormer [2] block

(c) Predict 3-DoF Pose Distribution

- Cross-Correlation (via FFT)



Data

We consider the following datasets:

Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	No	?

Data

We consider the following datasets:

Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	?

Google Maps, Bing Maps, Stratmap, DCGIS, MassGIS

Accurate georegistration? Problems:

- Invalid geo-pose of vehicle
- Invalid geo-registration of aerial images
- Hard to verify

Data – How to verify georegistration accuracy?

Is this registration accurate?

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Vehicle data: Ford AV dataset

Data – How to verify georegistration accuracy?

Is this registration accurate? ➔ yes

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Vehicle data: Ford AV dataset

Data – How to verify georegistration accuracy?

Is this registration accurate? ➔ no

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Vehicle data: Ford AV dataset

Data

We consider the following datasets:

Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	?

Accurate georegistration? Problems:

- Invalid geo-pose of vehicle
- Invalid geo-registration of aerial images
- Hard to verify

Data

We consider the following datasets:

Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	Manual labelling: Single frames

Accurate georegistration? Problems:
- ~~Invalid geo-pose of vehicle~~
- ~~Invalid geo-registration of aerial images~~
- ~~Hard to verify~~
- Can manually produce georegistration when lidar points are available

Data

We consider the following datasets:

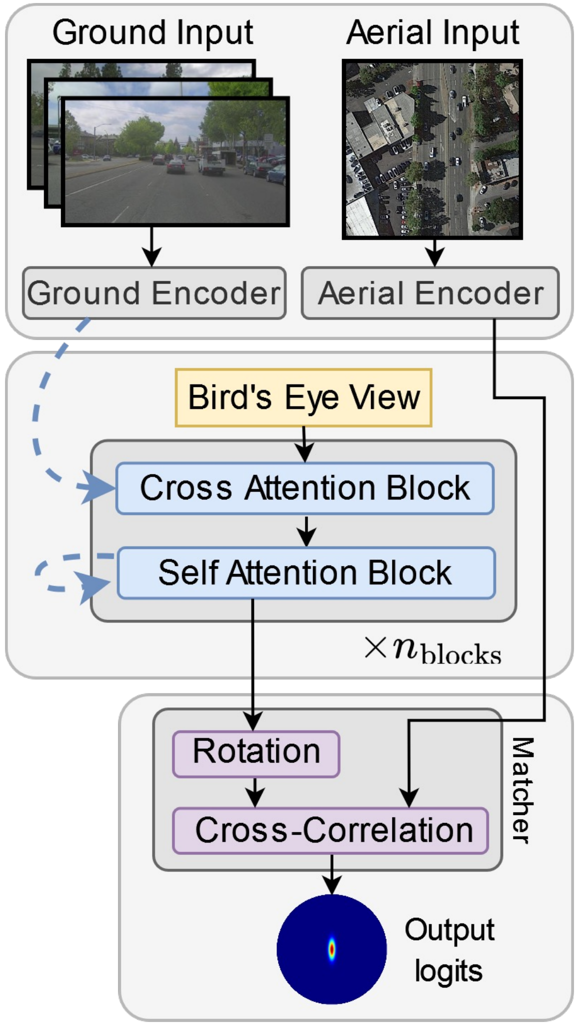
Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	Manual labelling: Trajectories

Accurate georegistration? Problems:

- ~~Invalid geo-pose of vehicle~~
- ~~Invalid geo-registration of aerial images~~
- ~~Hard to verify~~
- Can manually produce georegistration when lidar points and trajectories are available

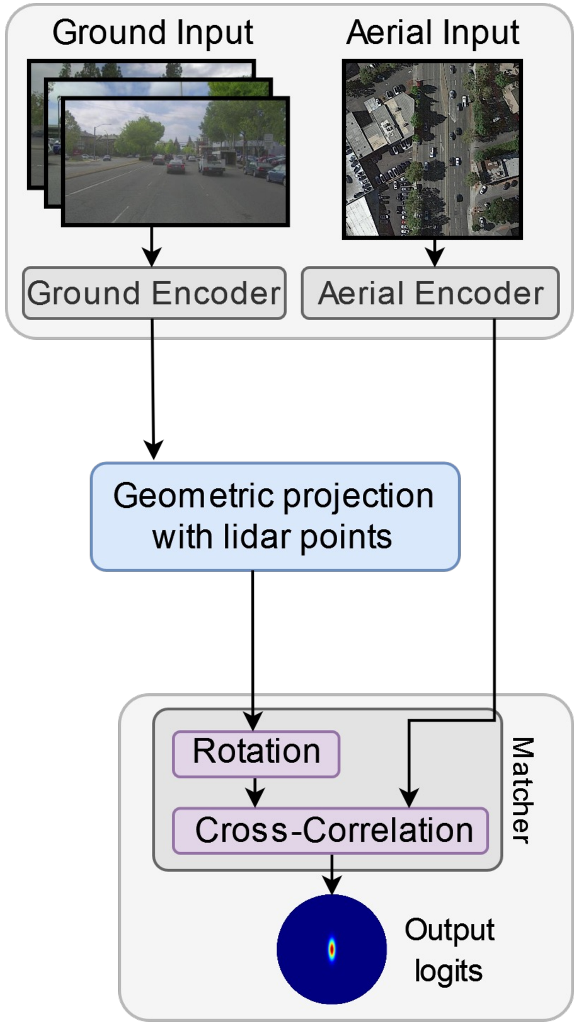
Pseudo-labels

Model:



Pseudo-labels

Model:



Steps:

1. Manually label subset of data
2. Train pseudo-label model on subset
3. Predict labels for all samples
4. Optimize using least squares
    a) Use inter-frame transforms with high confidence
    b) Use model predictions with low confidence

(5. Verify)

Data

We consider the following datasets:

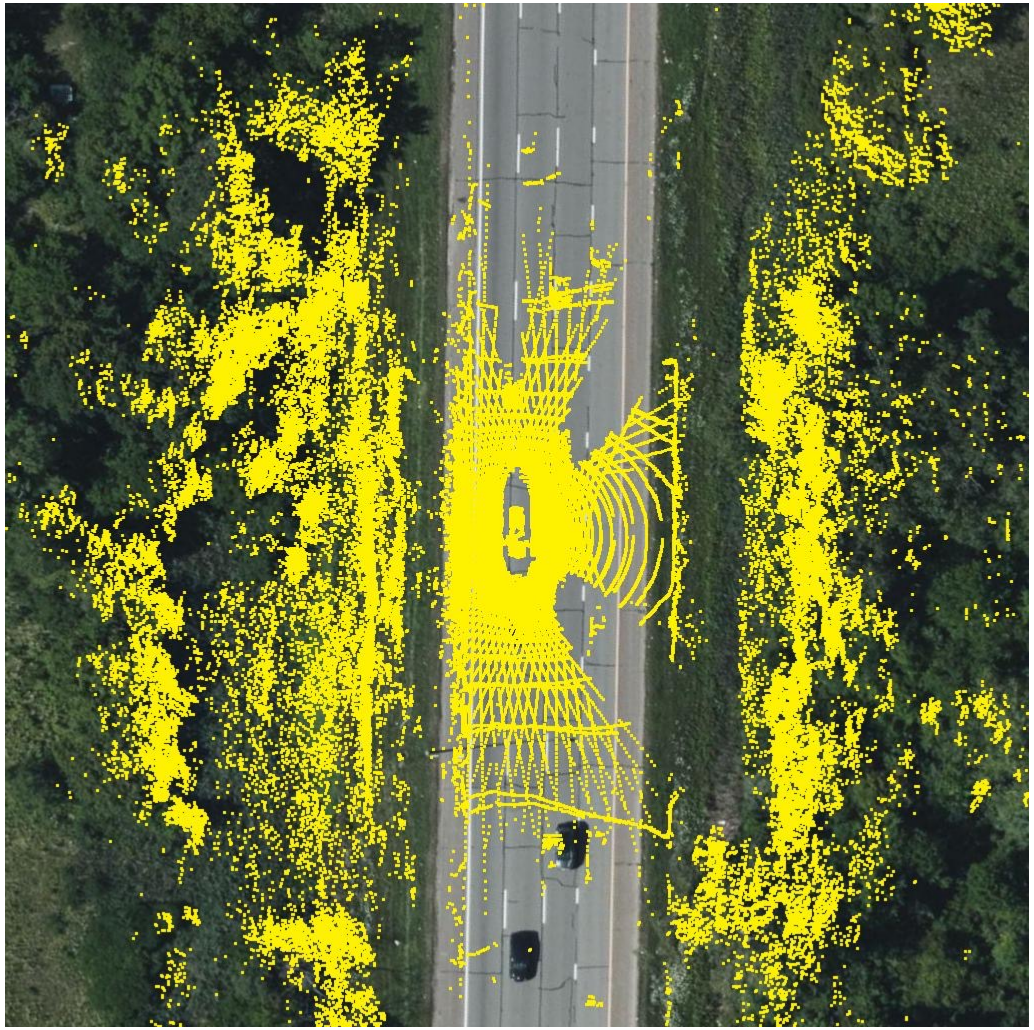
Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	Manual labelling: Trajectories

Data

We consider the following datasets:

Datasets from	Examples	Camera	Lidar	Trajectories	Aerial images	Accurate Georeg.
Large-Area CVGL	CVUSA, CVACT, VIGOR, …	Yes	No	No	Yes	?
Autonomous driving	KITTI, Ford AV, Nuscenes, …	Yes	Yes	Yes	(Yes)	Manual labelling Pseudo-labelling

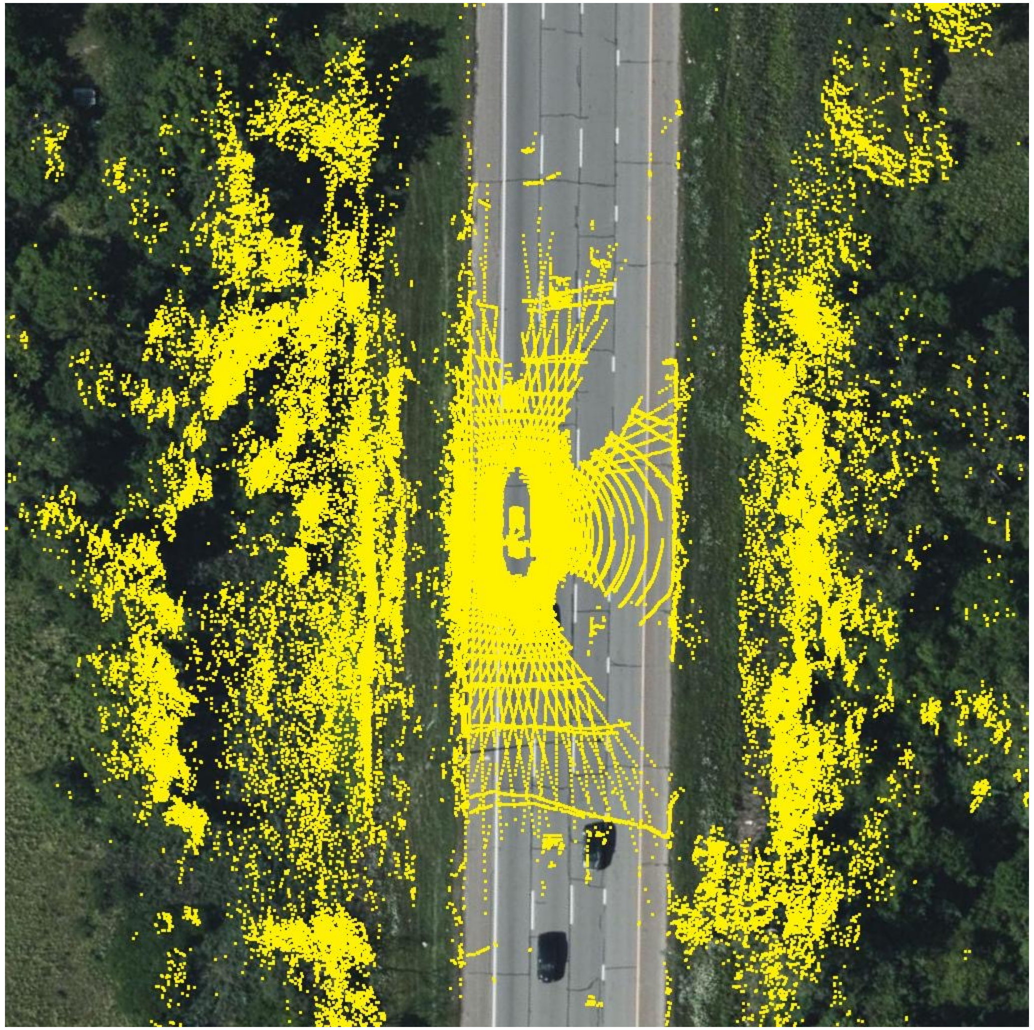
Data – Without Pseudo-labels



Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Data – With Pseudo-labels



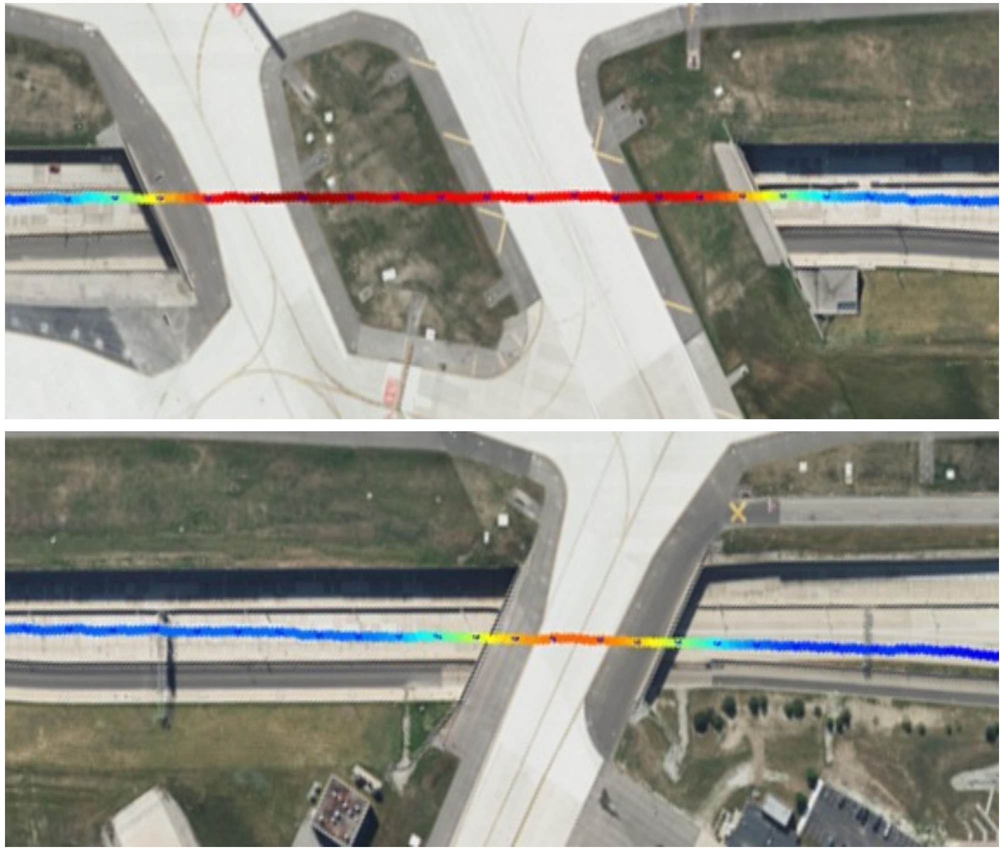
Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Data – Invalid data samples

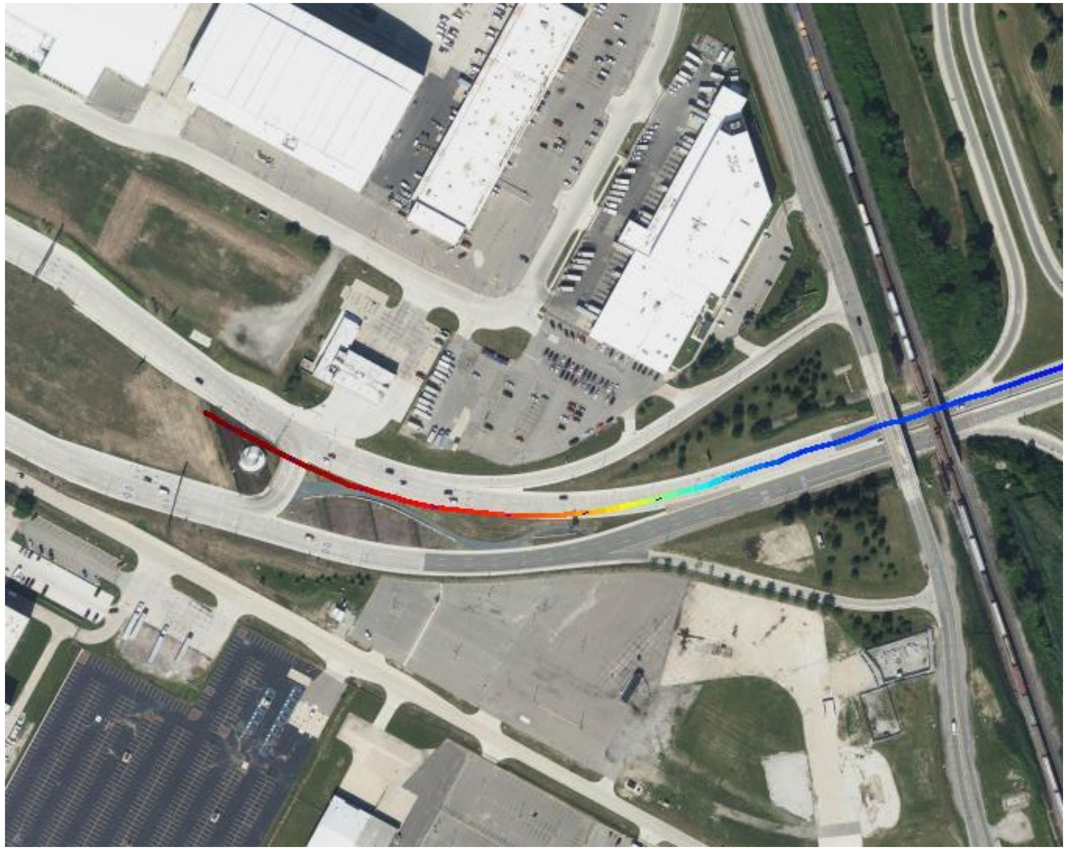
Remove data samples with low prediction confidence of pseudo-label model

Tunnel:



Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Out-of-date data:



Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Data

Dataset	Region	Year	Scenes	Frames ($\times 10^3$)	SD (sec)	Cams	Cells	Orthophoto providers
Argoverse V1 [11]	Miami	≤ 2019	53	12	22	9	71	Google Maps [3], Bing Maps [1]
	Pittsburgh	≤ 2019	60	10	17	9	55	Google Maps [3], Bing Maps [1]
Argoverse V2 [45]	Austin	≤ 2021	111	48	43	7	296	Google Maps [3], Bing Maps [1], Stratmap [5]
	Detroit	≤ 2021	256	91	36	7	569	Google Maps [3], Bing Maps [1]
	Miami	≤ 2021	703	245	34	7	811	Google Maps [3], Bing Maps [1]
	Palo Alto	≤ 2021	43	136	34	7	157	Google Maps [3], Bing Maps [1]
	Pittsburgh	≤ 2021	668	228	34	7	557	Google Maps [3], Bing Maps [1]
	Washington	≤ 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]

SD: Average scene duration. Data-frames are divided into disjoint cells with size 100m x 100m to measure aerial coverage.

Results

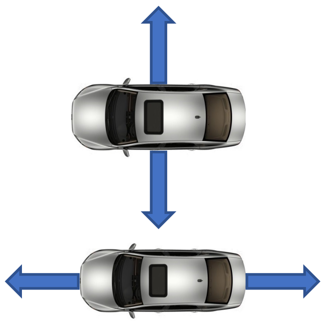
Recall on Ford AV (search region: ~28m, 20°):

	Cross-area	Cross-vehicle	Multi-camera	Log1						Log2					
				Lateral			Longitudinal			Lateral			Longitudinal		
				1.0m	3.0m	5.0m	1.0m	3.0m	5.0m	1.0m	3.0m	5.0m	1.0m	3.0m	5.0m
CVM-Net	✗	✗	✗	9.1	25.7	41.3	4.8	13.2	21.9	9.8	28.6	47.1	4.2	11.8	20.3
SAFA	✗	✗	✗	9.3	28.7	48.0	4.3	11.8	20.1	11.2	34.1	53.4	5.0	13.4	22.9
DSM	✗	✗	✗	12.0	35.3	53.7	4.3	12.5	21.4	8.5	24.9	37.6	3.9	12.2	21.4
VIGOR	✗	✗	✗	20.3	52.5	70.4	6.2	16.1	25.8	20.9	54.9	75.7	6.0	16.9	27.0
HighlyAccurate	✗	✗	✗	46.1	70.4	72.9	5.3	16.4	26.9	31.2	66.5	78.8	4.8	15.3	25.8
Ours	✗	✗	✗	**87.8**	**98.4**	**99.6**	**67.7**	**93.5**	**94.0**	**73.5**	**94.2**	**96.1**	**42.2**	**86.0**	**87.9**
Ours	✓	✓	✗	60.9	86.5	93.3	19.2	52.1	56.8	49.5	83.0	88.7	19.3	44.7	48.6
Ours	✗	✗	✓	96.3	99.6	99.6	76.0	95.3	96.0	88.0	99.9	100.0	58.9	93.3	93.6
Ours	✓	✓	✓	77.0	96.2	97.6	24.0	67.6	76.1	73.0	96.5	97.8	25.6	61.7	69.4

cross-area: *train/test data from non-overlapping regions*
cross-vehicle: *train/test data captured with different camera setup*

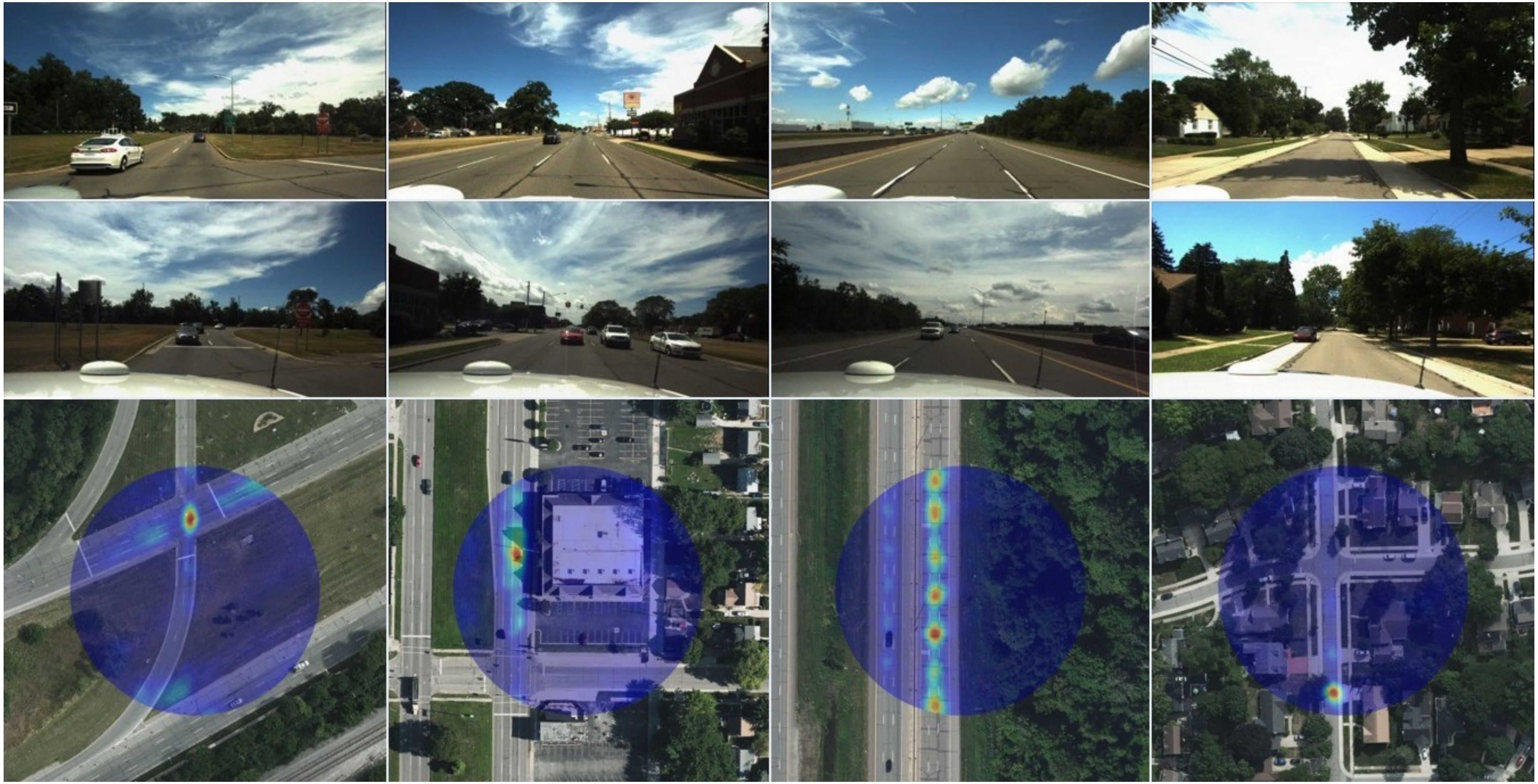
lateral recall

longitudinal recall



Results

Predictions on Ford AV (search region: ~28m, 20°):



Front Camera

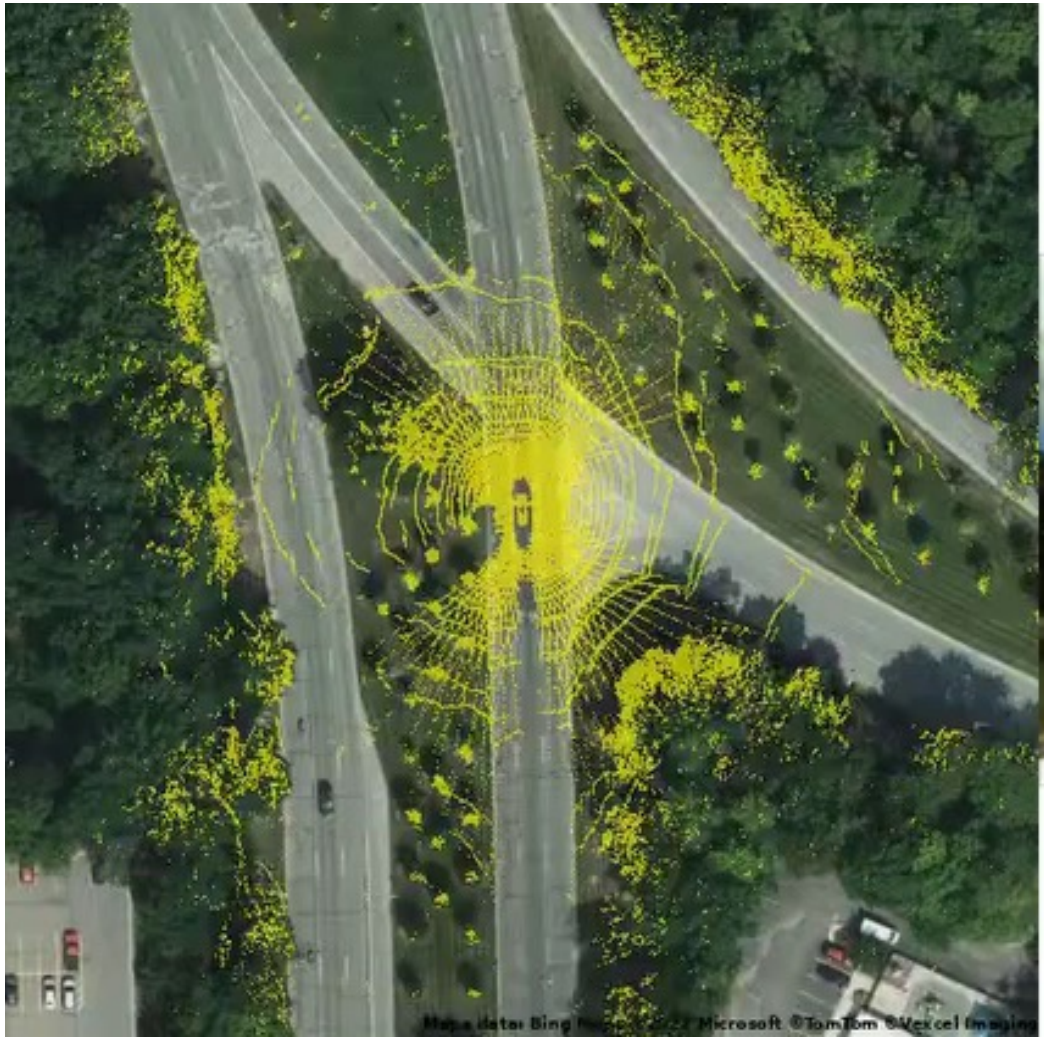
Back Camera

Model Prediction

Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Results

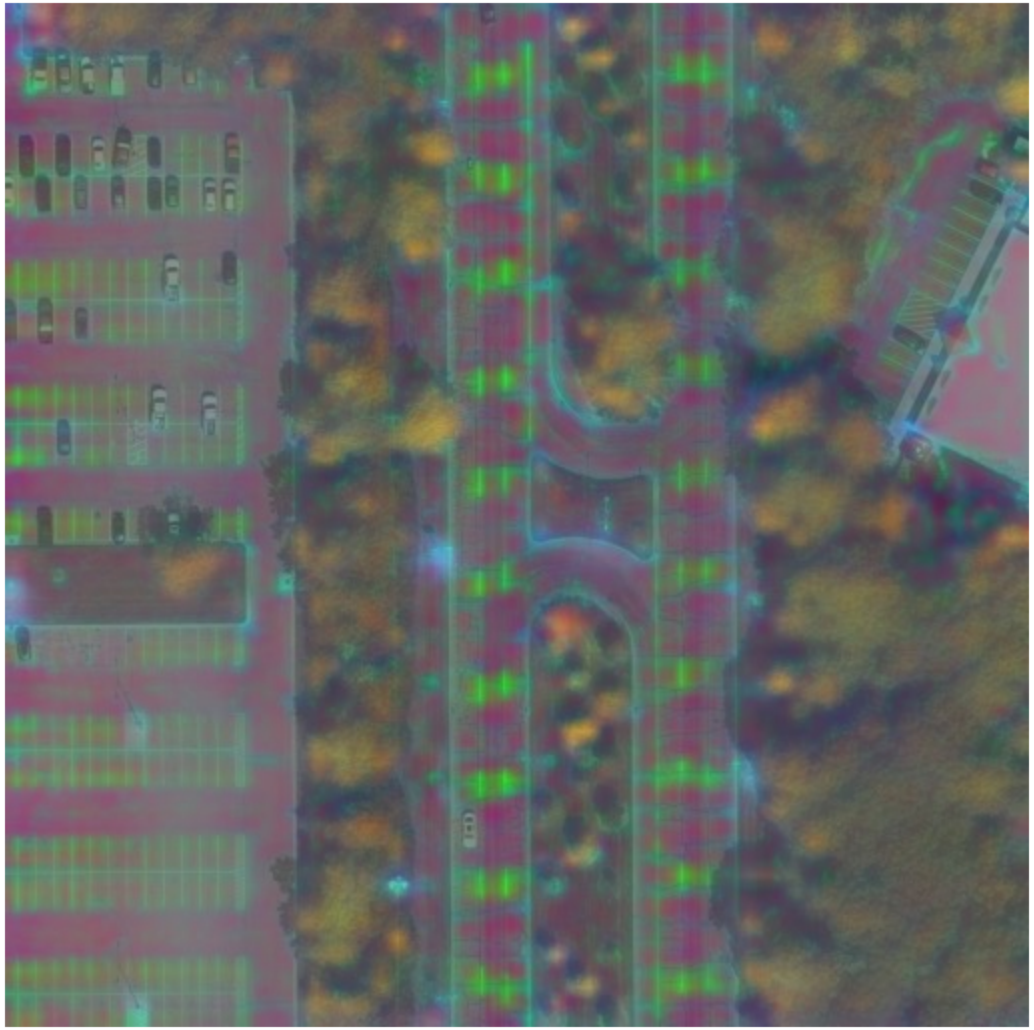
Tracker on Ford AV (10x speed):

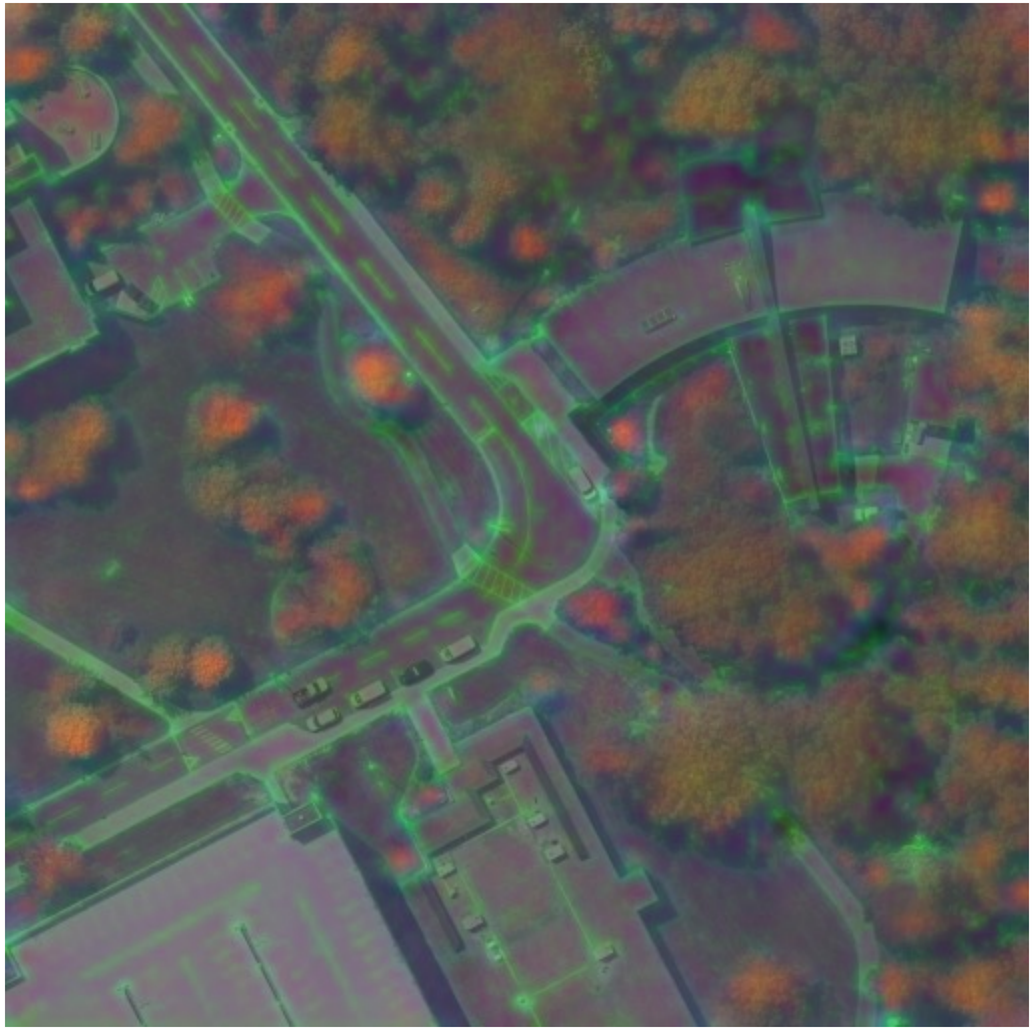


Aerial image (lidar scans shown for visualization only)

Front Camera

Feature Visualization – Ford AV





Map data: Bing Maps © 2022 TomTom, © Vexcel Imaging

Conclusion

- Related works:
  a) Low-level
  b) High-level semantic
  c) High-level end-to-end
- Novel model for vision-based metric CVGL
- State-of-the-art performance even in zero-shot setting
- Improved ground-truth for multiple datasets
  1. Pseudo-labels
  2. Automated data-pruning
- Code and ground-truth available online:

  https://fferflo.github.io/projects/vismetcvgl23

References

[1] Workman, Scott, Richard Souvenir, and Nathan Jacobs. "Wide-area image geolocalization with aerial reference imagery." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[2] Liu, Liu, and Hongdong Li. "Lending orientation to neural networks for cross-view geo-localization." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[3] Zhu, Sijie, Taojiannan Yang, and Chen Chen. "Vigor: Cross-view image geo-localization beyond one-to-one retrieval." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[4] Fervers, Florian, et al. "Uncertainty-aware Vision-based Metric Cross-view Geolocalization." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[5] Sarlin, Paul-Edouard, et al. "OrienterNet: Visual Localization in 2D Public Maps with Neural Matching." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[6] Lentsch, Ted, et al. "SliceMatch: Geometry-guided Aggregation for Cross-View Pose Estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[7] Noda, Masafumi, et al. "Vehicle ego-localization by matching in-vehicle camera images to an aerial image." Computer Vision–ACCV 2010 Workshops: ACCV 2010 International Workshops, Queenstown, New Zealand, November 8-9, 2010, Revised Selected Papers, Part II 10. Springer Berlin Heidelberg, 2011.
[8] Veronese, Lucas De Paula, et al. "Re-emission and satellite aerial maps applied to vehicle localization on urban environments." 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2015.
[9] Vysotska, Olga, and Cyrill Stachniss. "Improving SLAM by exploiting building information from publicly available maps and localization priors." PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science 85 (2017): 53-65.
[10] Kim, Jonghwi, and Jinwhan Kim. "Fusing lidar data and aerial imagery with perspective correction for precise localization in urban canyons." 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019.
[11] Brubaker, Marcus A., Andreas Geiger, and Raquel Urtasun. "Lost! leveraging the crowd for probabilistic visual self-localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2013.
[12] Floros, Georgios, Benito Van Der Zander, and Bastian Leibe. "Openstreetslam: Global vehicle localization using openstreetmaps." 2013 IEEE International Conference on Robotics and Automation. IEEE, 2013.
[13] Pink, Oliver. "Visual map matching and localization using a global feature map." 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2008.
[14] Javanmardi, Mahdi, et al. "Towards high-definition 3D urban mapping: Road feature-based registration of mobile mapping systems and aerial imagery." Remote Sensing 9.10 (2017): 975.
[15] Kümmerle, Rainer, et al. "Large scale graph-based SLAM using aerial images as prior information." Autonomous Robots 30 (2011): 25-39.
[16] Wang, Xipeng, Steve Vozar, and Edwin Olson. "Flag: Feature-based localization between air and ground." 2017 IEEE international conference on robotics and automation (ICRA). IEEE, 2017.
[17] Tang, Tim Yuqing, et al. "Rsl-net: Localising in satellite images from a radar on the ground." IEEE Robotics and Automation Letters 5.2 (2020): 1087-1094.
[18] Shi, Yujiao, and Hongdong Li. "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[19] Xia, Zimin, et al. "Cross-view matching for vehicle localization by learning geographically local representations." IEEE Robotics and Automation Letters 6.3 (2021): 5921-5928.
[20] Xia, Zimin, et al. "Visual cross-view metric localization with dense uncertainty estimates." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX. Cham: Springer Nature Switzerland, 2022.