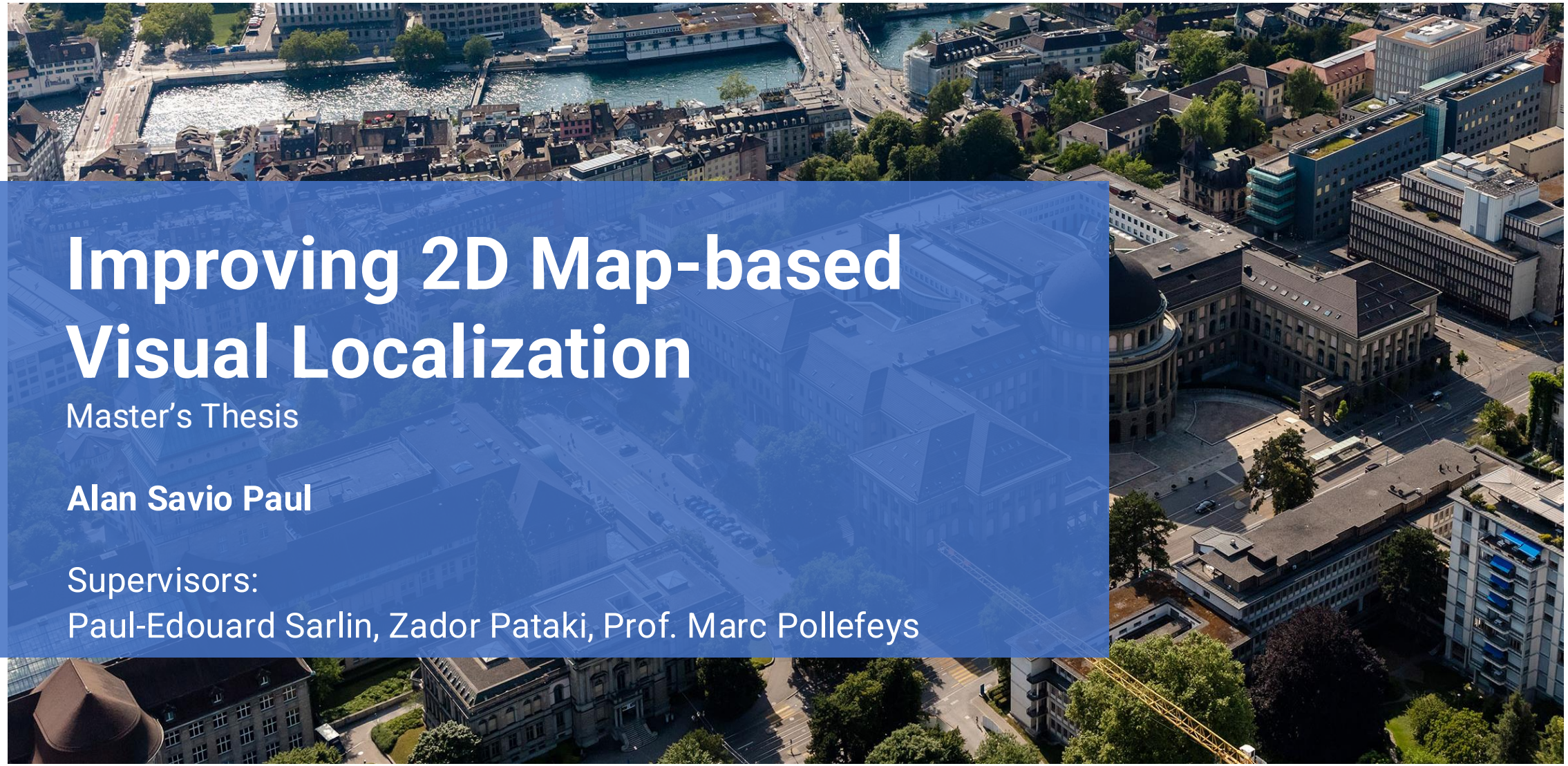
ETH zürich

Computer Vision and Geometry Lab



Improving 2D Map-based Visual Localization

Master's Thesis

Alan Savio Paul

Supervisors:
Paul-Edouard Sarlin, Zador Pataki, Prof. Marc Pollefeys

Why is this important?

- Robots and AR devices need to know where they are

Why is this important?

- Robots and AR devices need to know where they are
- 3D Maps are expensive to build and store at world scale
- 2D maps are free, lightweight, and contain sufficient information

Query Image

3D Map

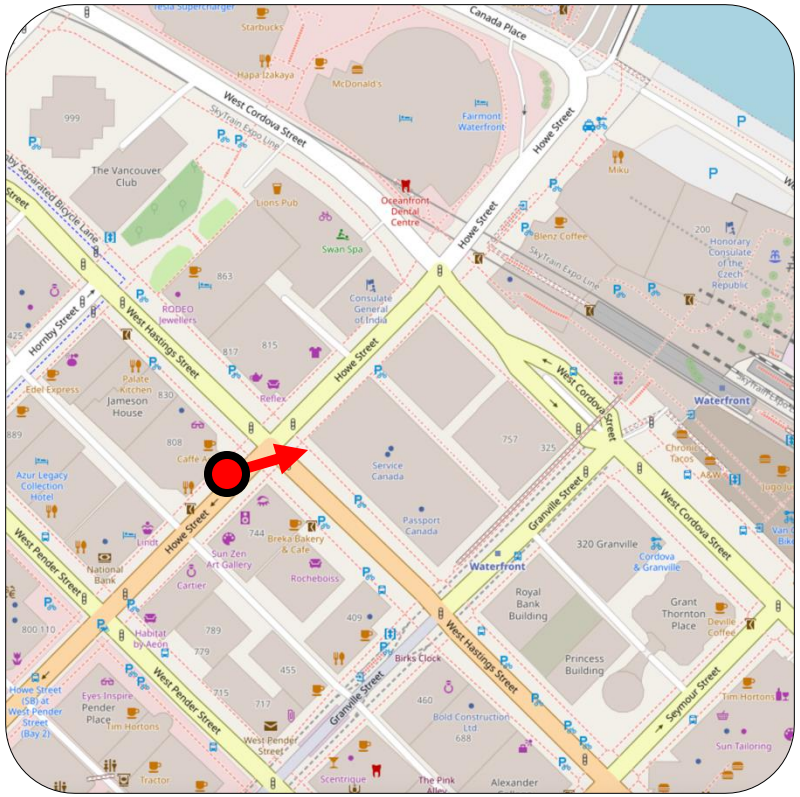
2D Map

Rich semantics in 2D Maps

Problem Setup

input RGB image

input map



output

3-DoF pose
(x, y, θ)

+ calibration: g , $\mathbf{K}$

Limitations of OrienterNet

- **Accuracy:** Does not utilize distant visual information.
- **Efficiency:** Cannot scale to larger distances and map sizes due to its high resolution (0.5mpp). Reducing resolution sacrifices accuracy.

Solving these would increase **practicality** of 2D-map based visual localization.

Thesis Goal

To improve the **accuracy**, **efficiency**, and thus, **practicality** of OrienterNet.

Ours

Ground Truth

Prediction

image

semantic map

Ours

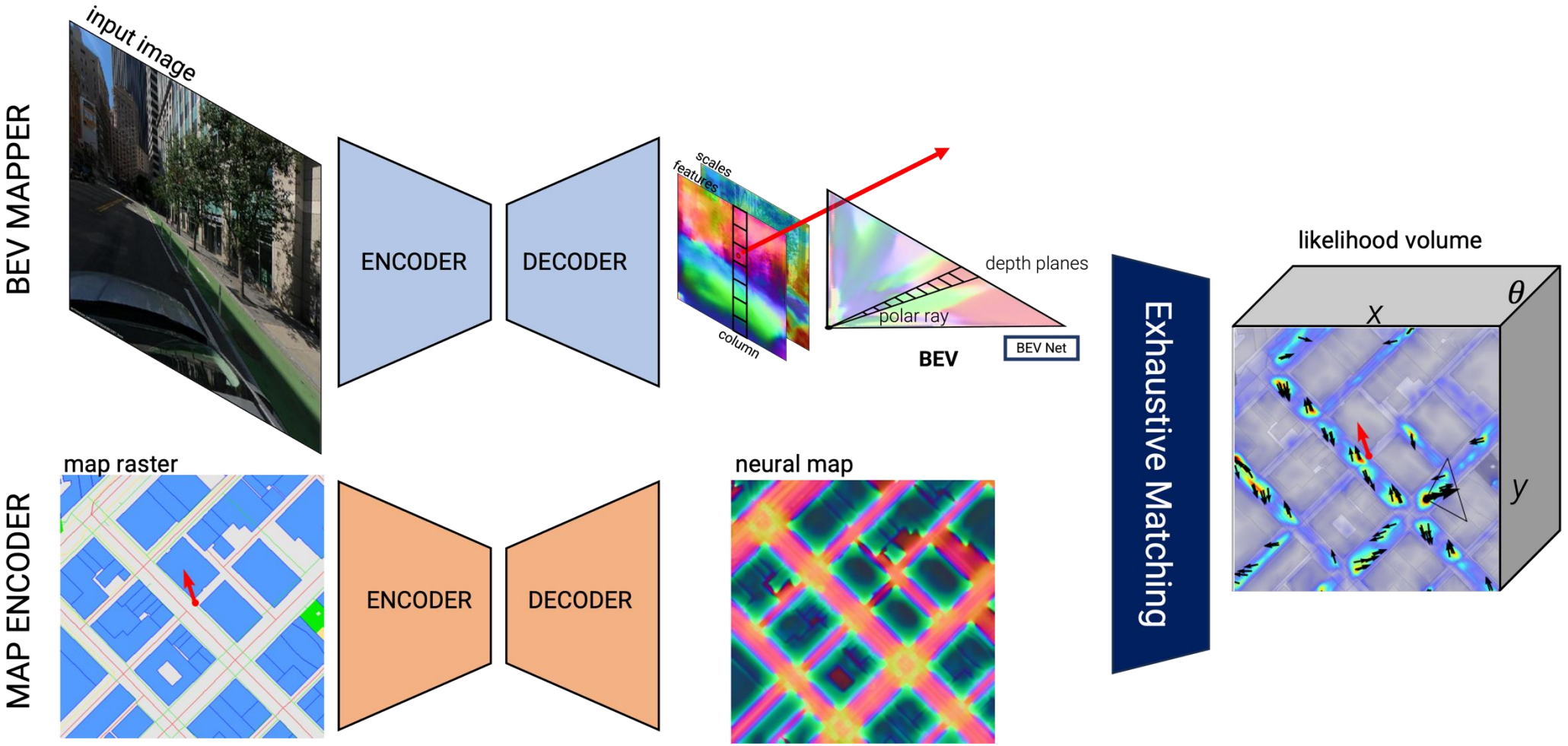
Ground Truth

Prediction

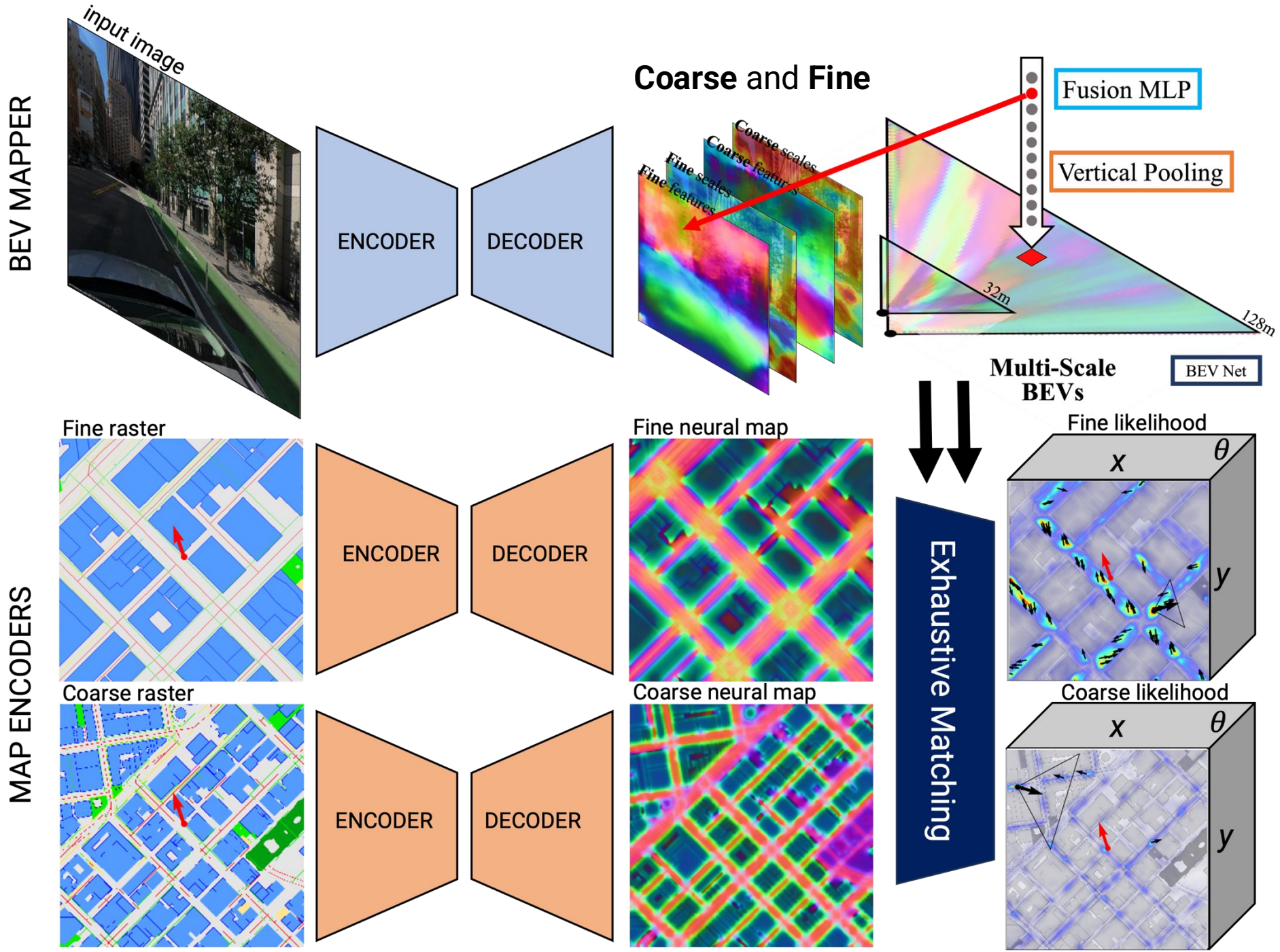
image

semantic map

Architecture: OrienterNet

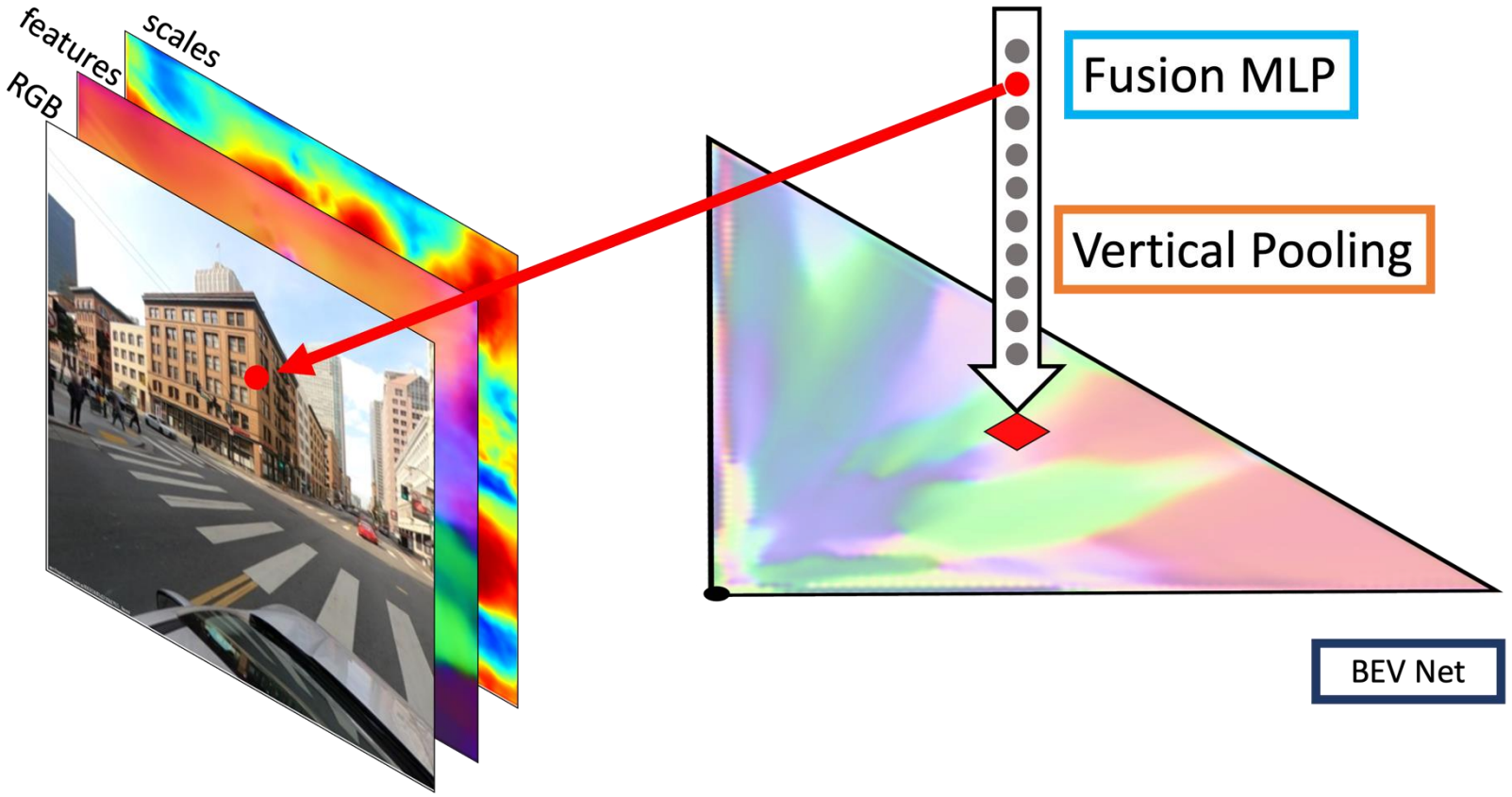


Architecture: Ours



BEV Mapper: Mapping Mechanism

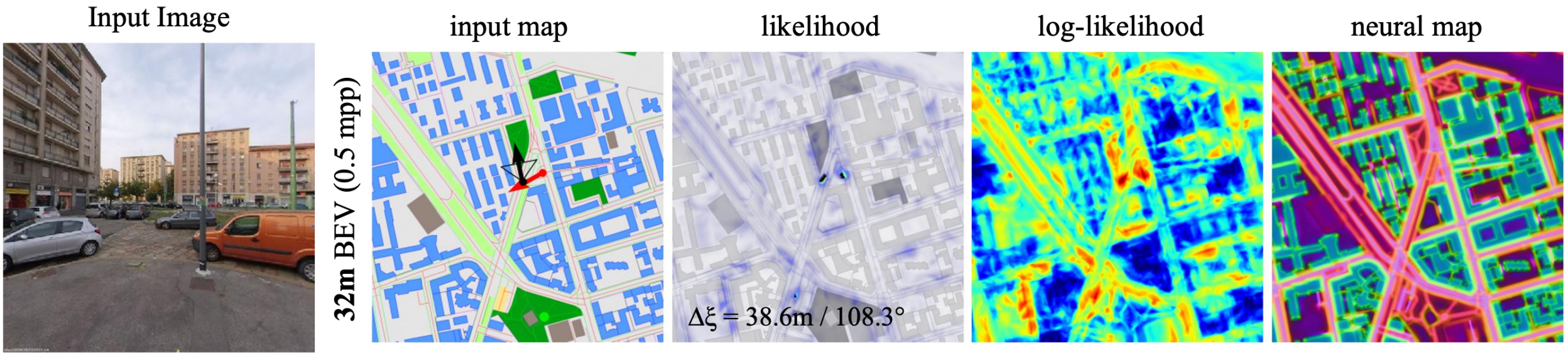
Inspired by **SNAP**, we use an Inverse Mapping strategy.



Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. **SNAP**: Self-Supervised Neural Maps for Visual Positioning and Semantic Understanding. In NeurIPS, 2023.

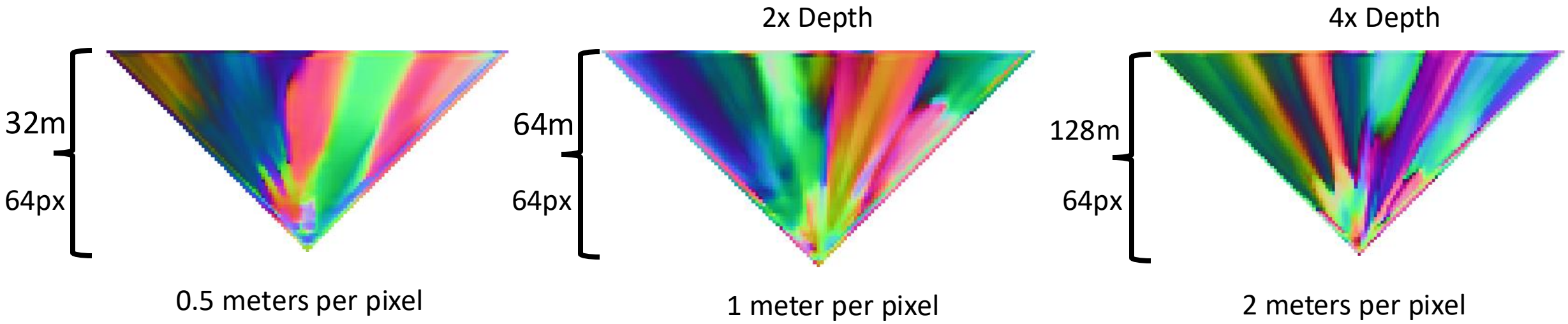
BEV Mapper: **Increasing Depth Range**

OrienterNet's BEV has a maximum depth of **32m**. This results in higher uncertainties.



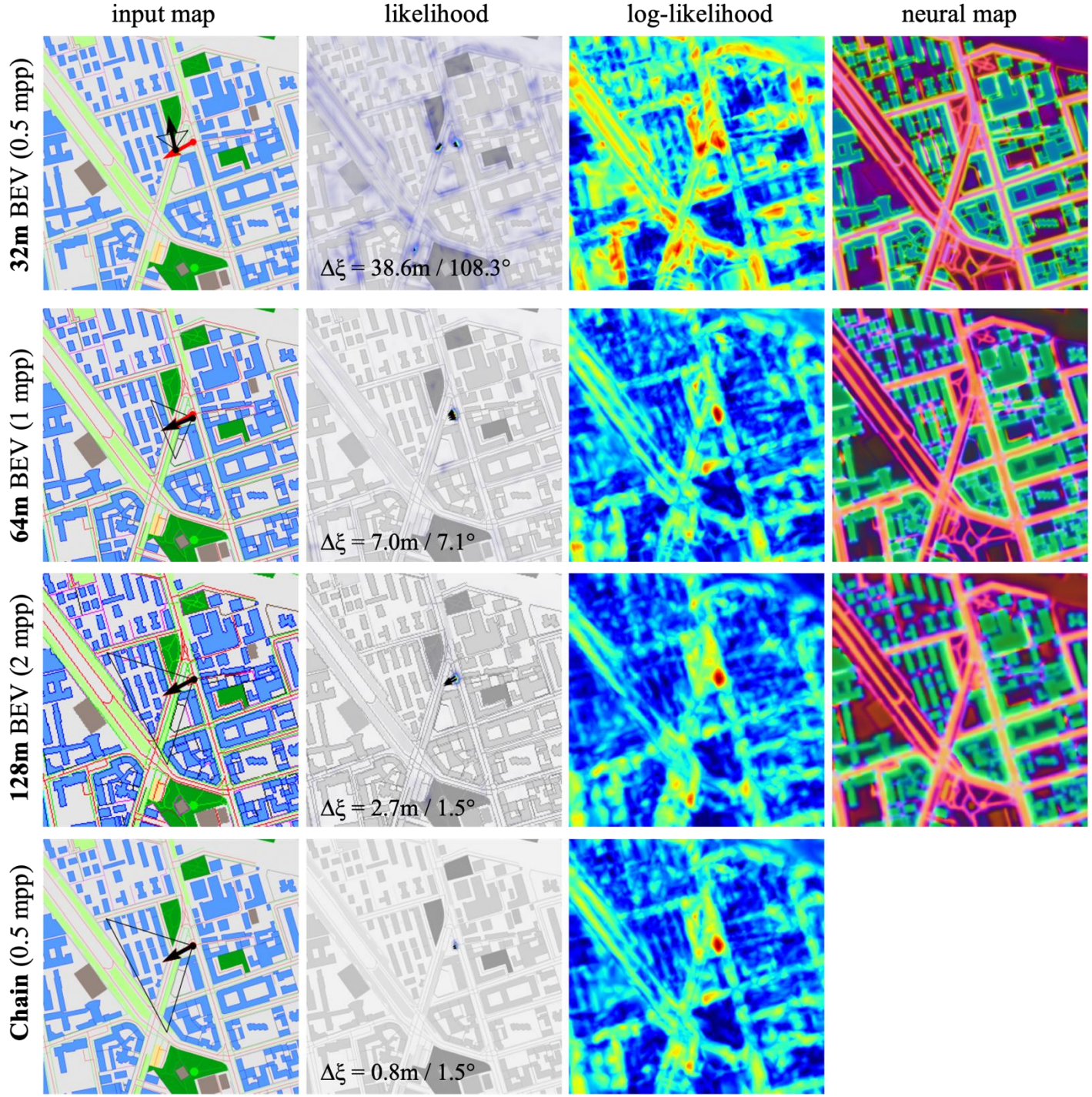
How can we enable Larger BEVs?

Multi-Scale BEVs



Larger depth, same computational cost

Input Image


input map
likelihood
log-likelihood
neural map
32m BEV (0.5 mpp)
Δξ = 38.6m / 108.3°
64m BEV (1 mpp)
Δξ = 7.0m / 7.1°
128m BEV (2 mpp)
Δξ = 2.7m / 1.5°
Chain (0.5 mpp)
Δξ = 0.8m / 1.5°

BEV Mapper: Increasing Depth Range

Search Radius: 256m

Model	Uncertainty	XY Mean Error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	Yaw Mean Error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
32m	0.36	144.07	1.2	4.19	11.72	19.99	23.29	25.95	64.49	4.19	9	16.9	31.92	39.19	43.59
64m	0.27	121.68	1.31	5.23	16.17	31.08	36.05	38.51	57.86	5.91	11.46	22.29	40.29	47.15	50.71
128m	0.21	113.77	1.2	3.61	12.14	31.61	39.25	42.8	53.25	6.44	12.3	24.18	43.85	51.23	55.15
Chain(32m, 64m)	0.24	112.61	2.2	6.33	19.52	34.33	39.25	42.39	53.61	6.8	13.13	25.07	45.21	51.75	55.1
Chain(64m,128m)	0.18	102.75	1.62	5.97	17.69	37.73	44.06	47.62	48.36	7.95	14.76	28.57	49.87	56.78	59.5
Chain(32m,128m)	0.25	103.48	1.52	6.07	19.05	35.85	42.44	45.94	49.01	**8.16**	**15.12**	28.68	49.29	55.99	59.55
Chain(32m, 64m, 128m)	0.2	**97.16**	**2.3**	**7.22**	**20.72**	**38.83**	**45.21**	**48.72**	**47.51**	7.95	14.81	**29.67**	**51.7**	**58.45**	**61.85**

Search Radius: 10m

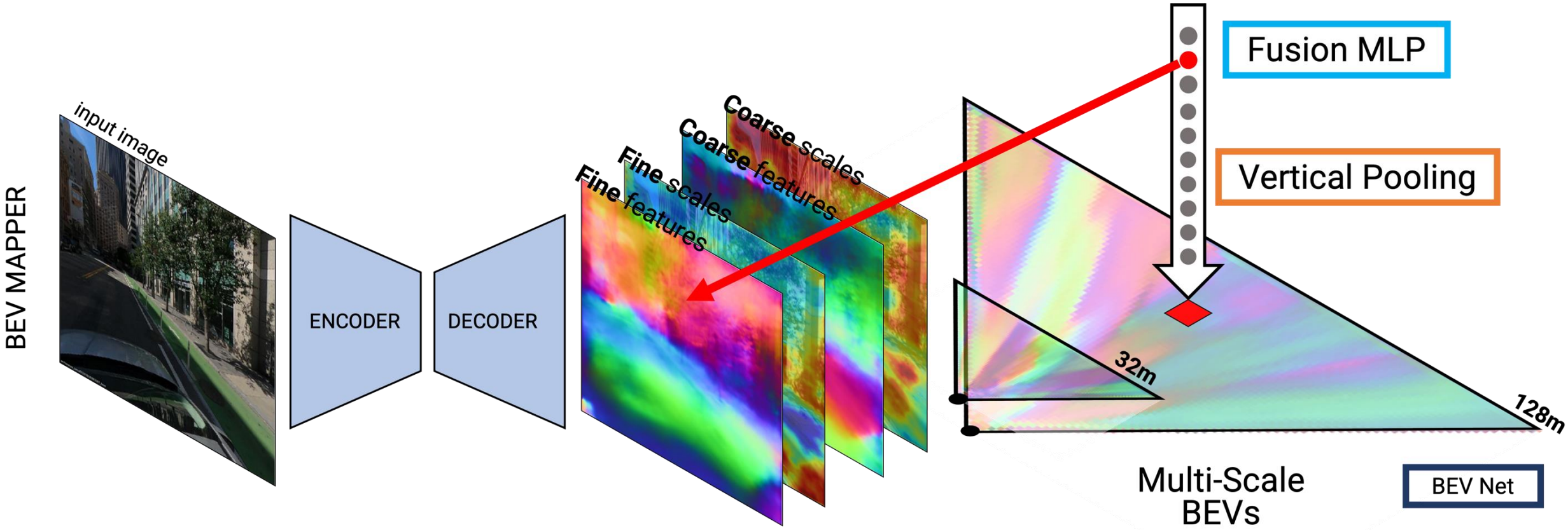
Model	Uncertainty	XY Mean Error (m)	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	Yaw Mean Error (°)	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
32m	0.27	4.87	2.51	10.2	29.04	60.28	86.81		18.56	10.05	18.94	36.21	68.5	83.73	88.23
64m	0.2	4.59	2.41	9.68	28.26	63.53	90.06		16.67	10.88	21.09	39.93	73.21	85.71	89.38
128m	0.15	4.7	1.73	5.91	20.62	65.1	90.63		11.4	12.24	22.87	42.86	76.03	88.23	92.94
Chain(32m, 64m)	0.21	4.46	**3.24**	**11.67**	32.34	65.1	89.74		14.51	11.62	22.71	43.38	76.3	88.17	91.31
Chain(64m,128m)	0.15	4.52	2.41	9.58	28.78	64.15	90.32		11.81	12.3	23.5	44.22	78.02	89.17	92.67
Chain(32m,128m)	0.21	4.44	2.51	10.41	31.71	65.1	89.95		10.94	12.24	23.55	44.69	79.7	90.95	93.67
Chain(32m, 64m, 128m)	0.18	**4.32**	3.14	11.04	**33.49**	**66.25**	90.53		**10.17**	**12.87**	**24.44**	**46.52**	**80.74**	**91.42**	**94.09**

Higher Depth Range + Coarser Resolution → Better Retrieval, Worse Fine Accuracy

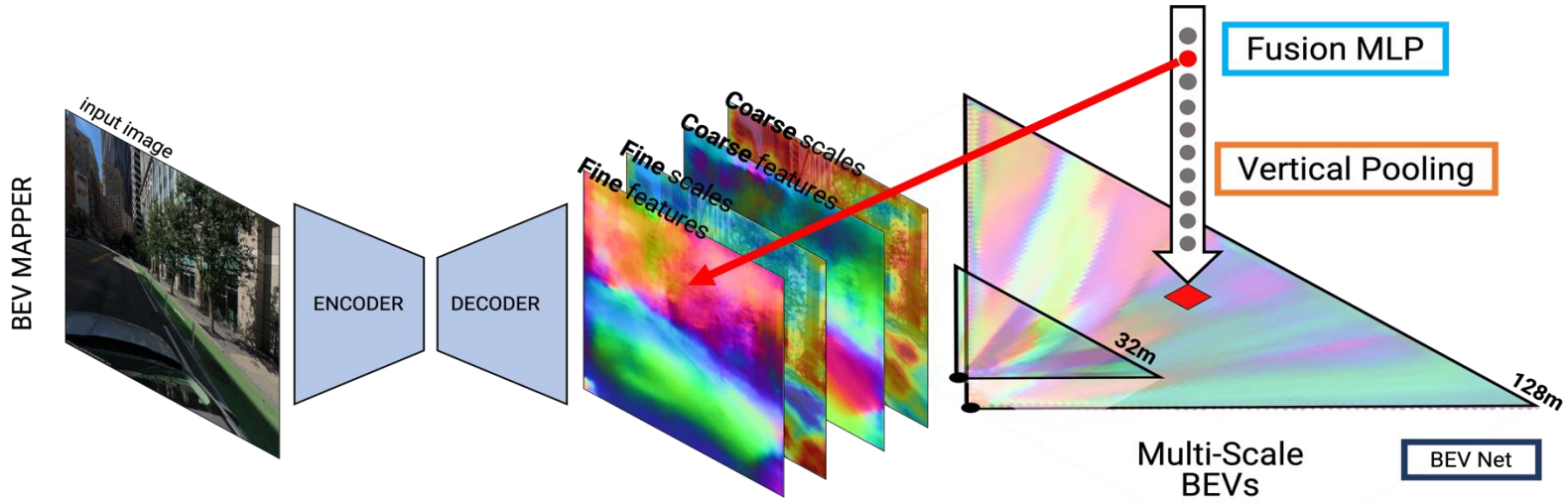
However, Chaining Fine + Coarse → Best of Both Worlds

BEV Mapper

We predict two BEVs: **fine** (depth: 32m) and **coarse** (depth: 128m)



BEV Mapper

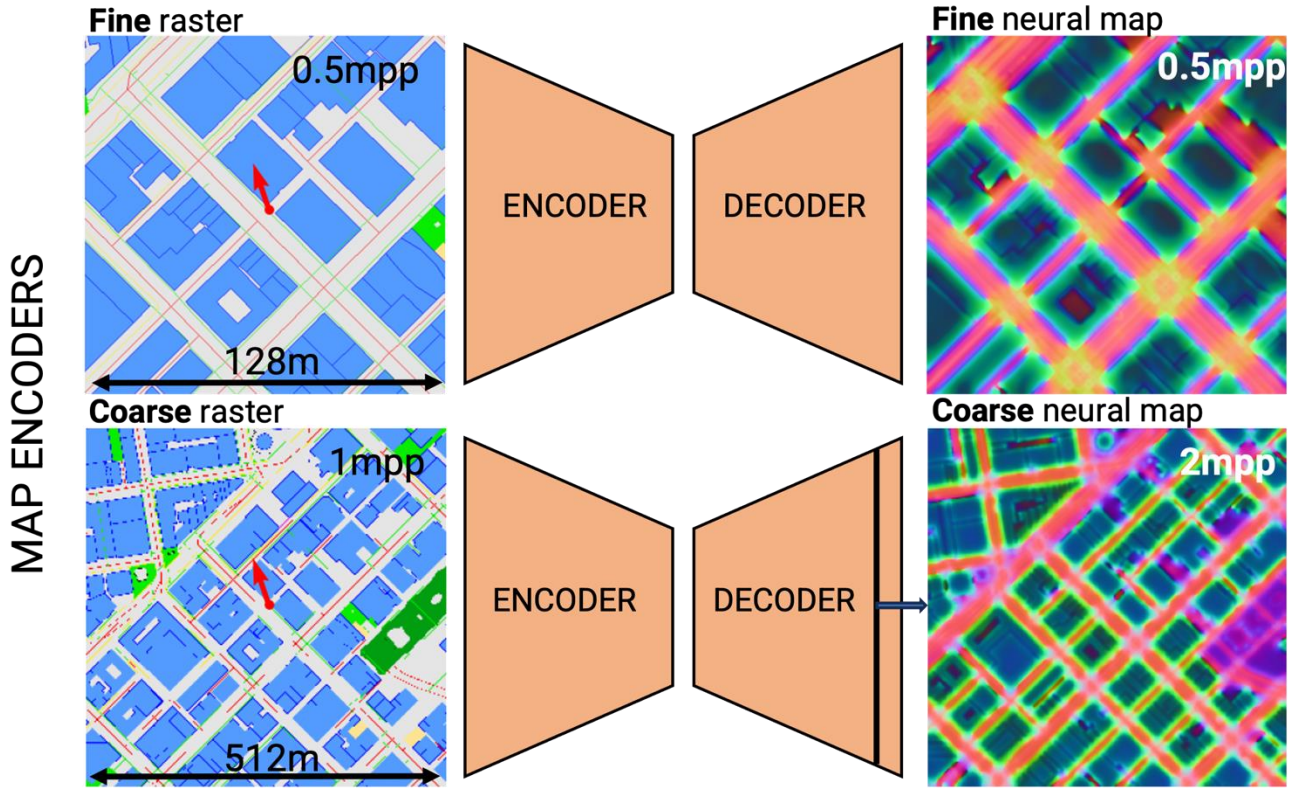


Fine branch can learn detailed occupancy of nearby objects (eg. poles, building corners)

Coarse branch can learn occupancy of both near and far objects (eg. presence of building, river, park).

Neural Map Encoding

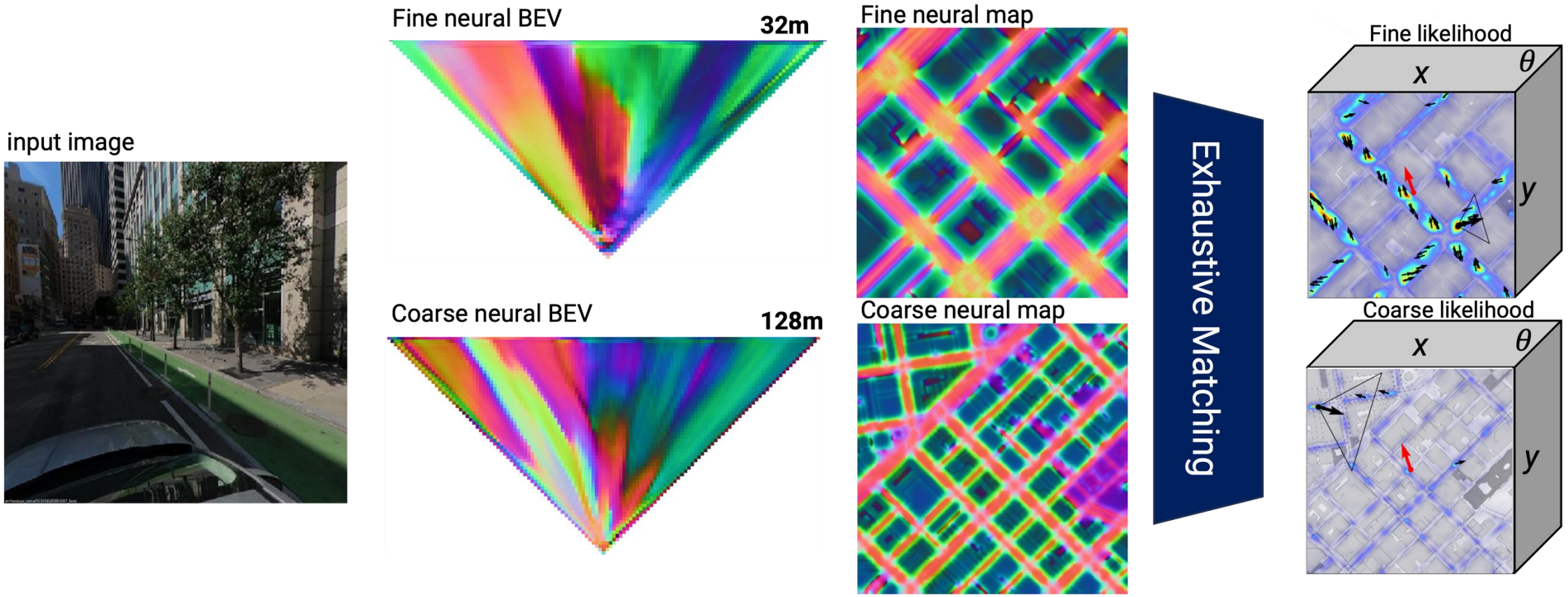
We predict two Neural Maps: **fine** and **coarse**.



During training, both neural maps are 256px x 256px

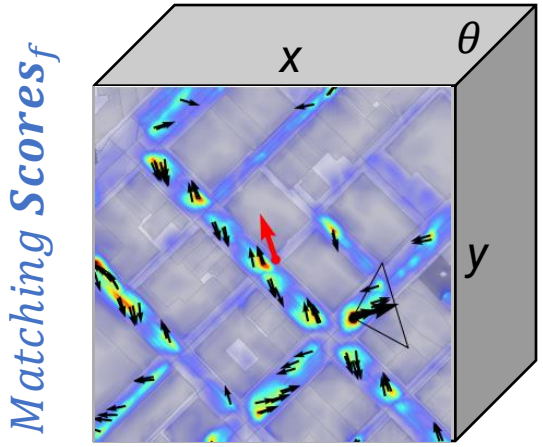
Pose Estimation

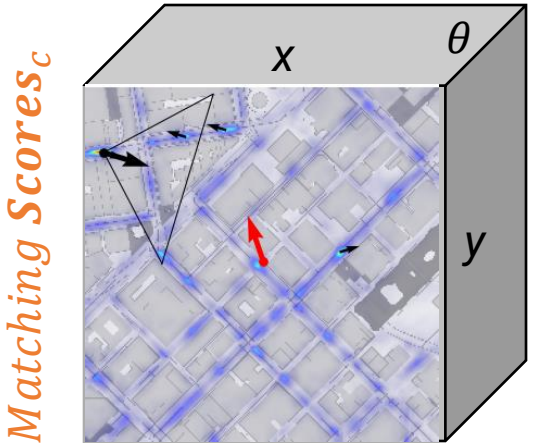
- Given a **Neural BEV** and a **Neural Map**, we exhaustively match them to obtain a **likelihood volume**.



Training Loss

Results in 5% increase in R@20m over trivial sum





We maximise the likelihood of GT pose in each branch

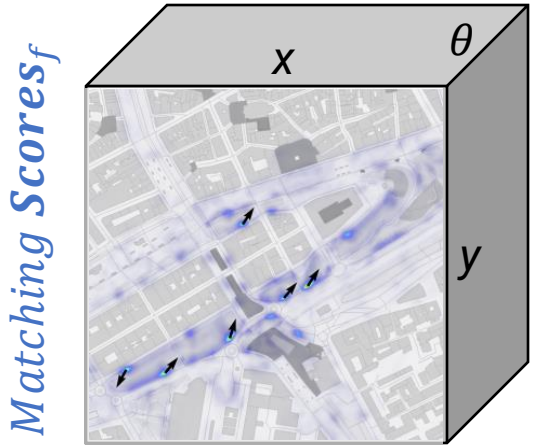
The two losses are weighted using a predicted task uncertainty

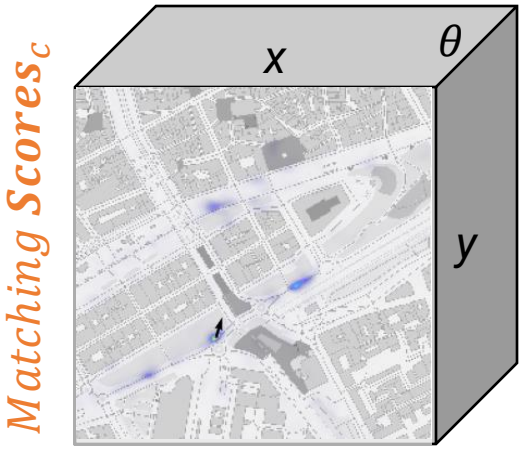
$$\boldsymbol{P_f} = \textbf{Softmax}\left(\frac{1}{\sigma_f^2}.\text{Scores}_f\right), \boldsymbol{P_c} = \textbf{Softmax}\left(\frac{1}{\sigma_c^2}.\text{Scores}_c\right)$$

$$\textbf{Loss} = -\log \boldsymbol{P_f}[pose_{gt}] - \log \boldsymbol{P_c}[pose_{gt}] + \log \sigma_f + \log \sigma_c$$

Pose Estimation: Inference

Query Image





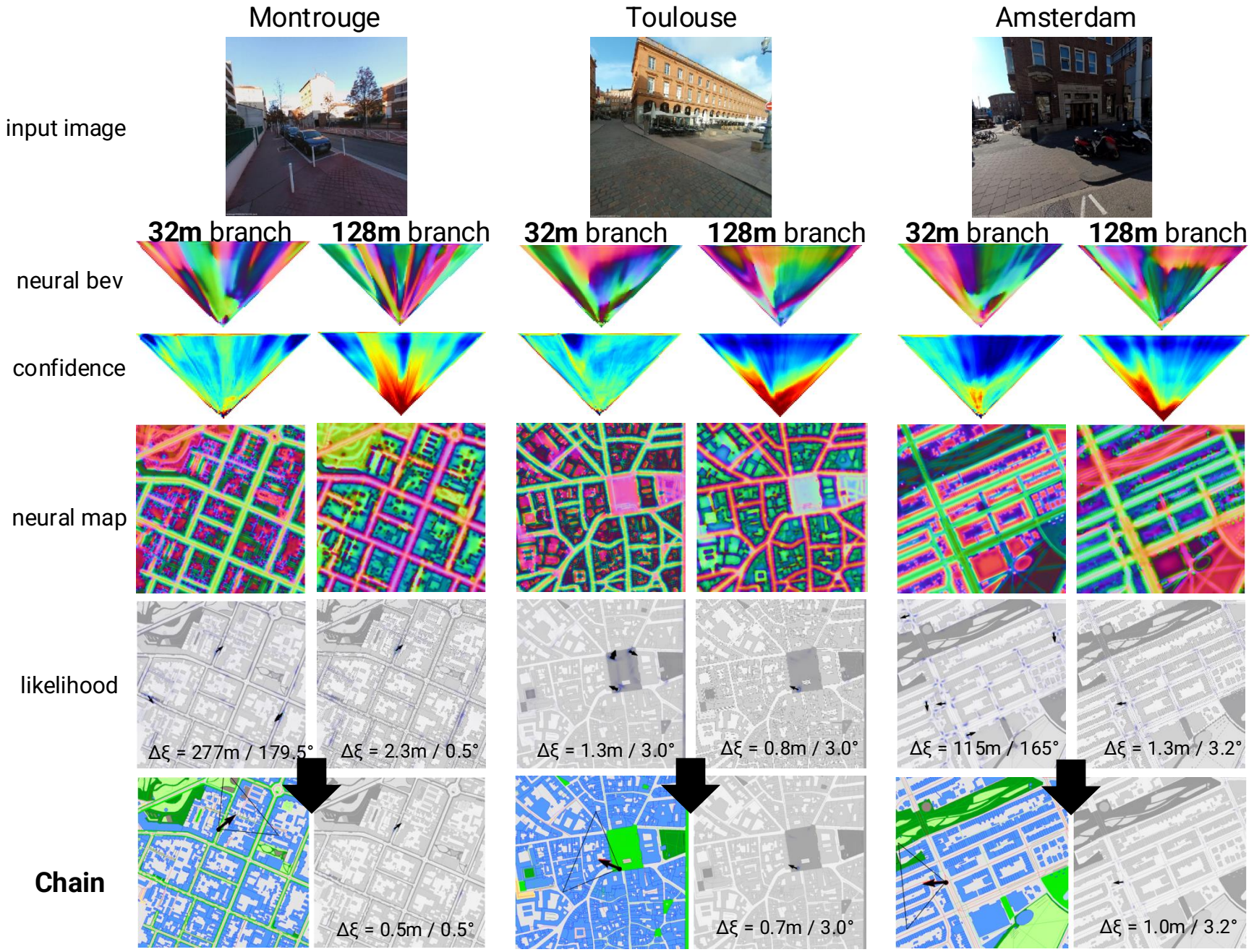
Joint Likelihood: We multiply both likelihood volumes.

Both volumes cover same areas. The coarser volume has fewer pixels, so we bilinearly upsample it.

$$\boldsymbol{P_f} = \textbf{Softmax}\left(\frac{1}{\sigma_f^2}.\text{Scores}_f\right), \boldsymbol{P_c} = \textbf{Softmax}\left(\frac{1}{\sigma_c^2}.\mathcal{U}(\text{Scores}_c)\right)$$

$$\boldsymbol{P_{chain}} = \boldsymbol{P_f}[pose_{gt}] * \boldsymbol{P_c}[pose_{gt}]$$

Pose Estimation



Results: Comparing with Single Branch Models

Search Radius: 256m

Model	Uncertainty	XY Mean Error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	Yaw Mean Error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
32m	0.36	144.07	1.2	4.19	11.72	19.99	23.29	25.95	64.49	4.19	9	16.9	31.92	39.19	43.59
64m	0.27	121.68	1.31	5.23	16.17	31.08	36.05	38.51	57.86	5.91	11.46	22.29	40.29	47.15	50.71
128m	0.21	113.77	1.2	3.61	12.14	31.61	39.25	42.8	53.25	6.44	12.3	24.18	43.85	51.23	55.15
Chain(32m, 64m)	0.24	112.61	2.2	6.33	19.52	34.33	39.25	42.39	53.61	6.8	13.13	25.07	45.21	51.75	55.1
Chain(64m,128m)	0.18	102.75	1.62	5.97	17.69	37.73	44.06	47.62	48.36	7.95	14.76	28.57	49.87	56.78	59.5
Chain(32m,128m)	0.25	103.48	1.52	6.07	19.05	35.85	42.44	45.94	49.01	**8.16**	**15.12**	28.68	49.29	55.99	59.55
Chain(32m, 64m, 128m)	0.2	**97.16**	2.3	7.22	20.72	38.83	45.21	48.72	47.51	7.95	14.81	**29.67**	51.7	58.45	61.85
Multi-Scale (32m branch)	0.31	130.6	1.73	5.08	16.12	25.64	29.57	32.76	59.87	4.4	8.95	18.11	35.01	43.43	48.61
Multi-Scale (128m branch)	0.18	106	1.15	4.87	17.16	35.95	43.49	47.25	46.78	7.27	14.08	27	49.56	56.93	60.65
Multi-Scale	0.22	97.49	**2.72**	**8.53**	**22.92**	**39.77**	**46.62**	**50.03**	**43.64**	**8.16**	**15.12**	29.15	**53.17**	**59.97**	**63.63**

Search Radius: 10m

Model	Uncertainty	XY Mean Error (m)	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	Yaw Mean Error (°)	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
32m	0.27	4.87	2.51	10.2	29.04	60.28	86.81		18.56	10.05	18.94	36.21	68.5	83.73	88.23
64m	0.2	4.59	2.41	9.68	28.26	63.53	90.06		16.67	10.88	21.09	39.93	73.21	85.71	89.38
128m	0.15	4.7	1.73	5.91	20.62	65.1	90.63		11.4	12.24	22.87	42.86	76.03	88.23	92.94
Chain(32m, 64m)	0.21	4.46	3.24	**11.67**	32.34	65.1	89.74		14.51	11.62	22.71	43.38	76.3	88.17	91.31
Chain(64m,128m)	0.15	4.52	2.41	9.58	28.78	64.15	90.32		11.81	12.3	23.5	44.22	78.02	89.17	92.67
Chain(32m,128m)	0.21	4.44	2.51	10.41	31.71	65.1	89.95		10.94	12.24	23.55	44.69	79.7	90.95	93.67
Chain(32m, 64m, 128m)	0.18	4.32	3.14	11.04	33.49	66.25	90.53		**10.17**	**12.87**	**24.44**	**46.52**	**80.74**	**91.42**	**94.09**
Multi-Scale (32m branch)	0.25	4.6	3.19	10.99	32.76	62.79	89.01		16.43	9.79	19	36.16	70.33	84.51	89.69
Multi-Scale (128m branch)	0.13	4.43	1.67	7.74	28.05	66.09	**91.78**		11.41	12.19	23.23	43.22	78.49	89.01	93.2
Multi-Scale	0.19	**4.25**	**3.72**	11.62	**34.69**	**66.41**	91.31		10.92	12.66	23.81	44.27	80.22	90.69	93.72

Results: Comparing with OrienterNet

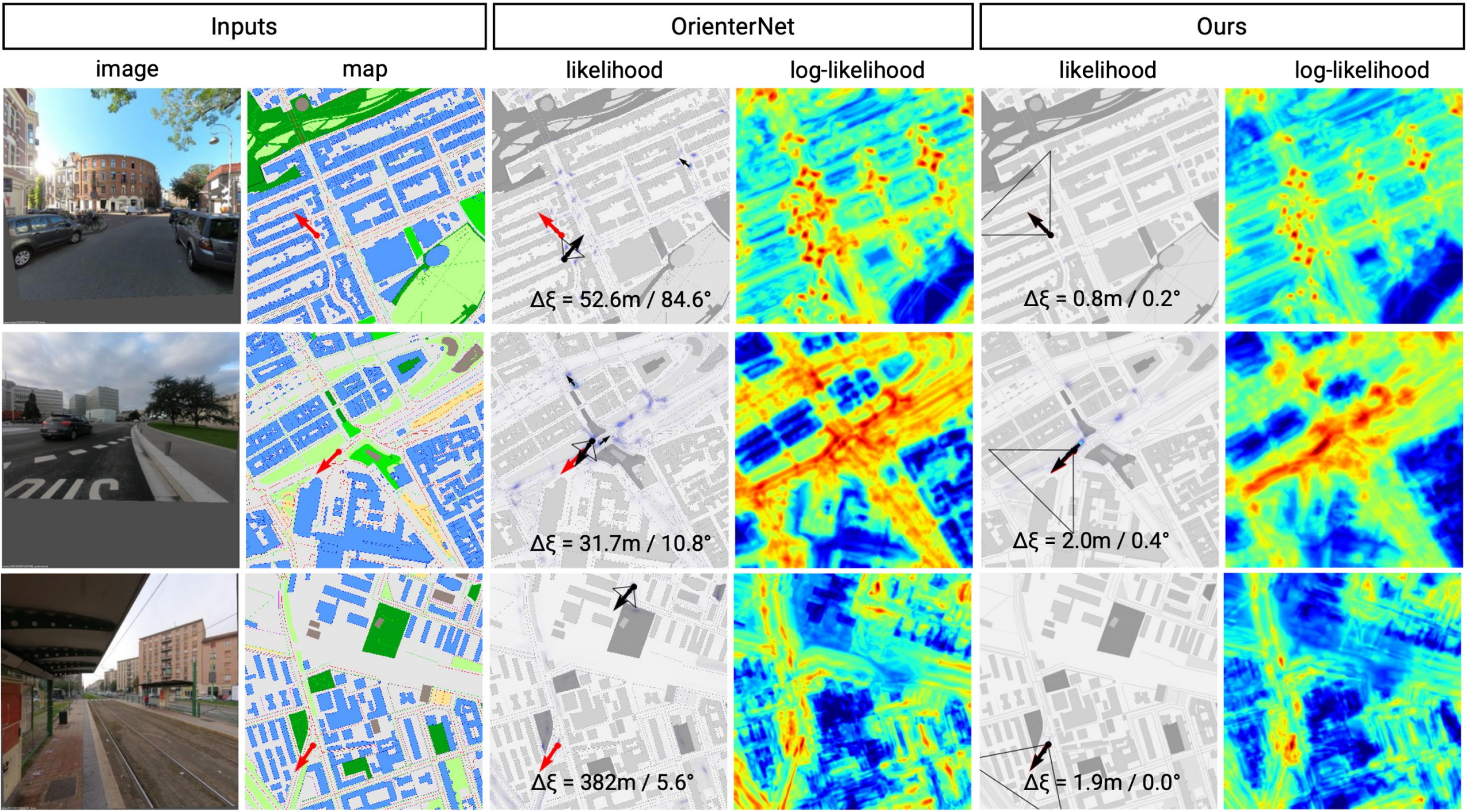
Search Radius: 256m

experiment	uncertainty	xy mean error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw mean error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
OrienterNet	-	109.58	2.83	9	22.92	34.38	38.41	41.23	52.79	6.54	12.3	23.29	42.86	50.81	54.74
Ours (32m branch)	0.33	109.5	2.15	8.11	20.83	32.86	37.57	40.87	55.6	6.12	11.25	22.19	41.81	49.92	53.43
Ours (128m branch)	0.18	78.94	2.56	7.8	23.5	45.79	54.42	57.51	38.16	9.21	**17.53**	33.96	57.98	65.67	69.13
Ours	0.22	**71.14**	**3.35**	**11.67**	**29.98**	**48.82**	**56.62**	**60.07**	**37.42**	**9.47**	**17.53**	**34.9**	**60.02**	**67.66**	**70.02**

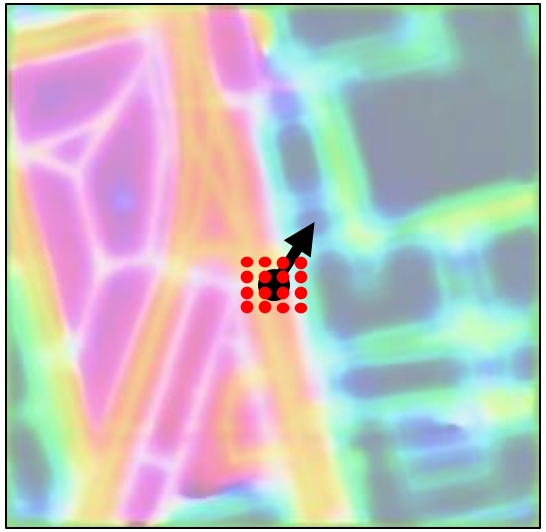
Search Radius: 10m

experiment	uncertainty	xy error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
OrienterNet		4.06	**5.02**	15.65	39.72	69.07	91.05		12.58	11.25	21.04	40.03	75.82	88.8	92.67
Ours (32m branch)	0.26	4.11	3.87	14.91	37.47	68.29	91		12.98	11.2	20.36	39.87	74.1	88.96	92.67
Ours (128m branch)	0.13	3.95	3.77	11.36	34.64	70.96	**93.56**		**8.11**	13.24	25.01	46.89	81.79	92.78	**95.55**
Ours	0.2	**3.83**	4.24	**16.06**	**40.55**	**71.27**	92.83		8.5	**13.55**	**26.01**	**49.5**	**83.73**	**93.41**	95.5

Our model learns to optimally leverage coarse and fine BEVs and maps to outperform OrienterNet by over 20% Recall at 20m in the larger map.


Inputs
OrienterNet
Ours
image
map
likelihood
log-likelihood
likelihood
log-likelihood
Δξ = 52.6m / 84.6°
Δξ = 0.8m / 0.2°
Δξ = 31.7m / 10.8°
Δξ = 2.0m / 0.4°
Δξ = 382m / 5.6°
Δξ = 1.9m / 0.0°

Pose Refinement



Exhaustively search through a **pose grid** of size **2m x 2m x 14°** and resolution **0.1m x 0.5°**, centered at argmax.

Pose Refinement

Grid Size: **2m x 2m x 14°**
Grid resolution: **0.1m x 0.5°**

Search Radius: 256m

experiment	uncertainty	xy mean error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw mean error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
OrienterNet	-	109.58	2.83	9	22.92	34.38	38.41	41.23	52.79	6.54	12.3	23.29	42.86	50.81	54.74
Ours (32m branch)	0.33	109.5	2.15	8.11	20.83	32.86	37.57	40.87	55.6	6.12	11.25	22.19	41.81	49.92	53.43
Ours (128m branch)	0.18	78.94	2.56	7.8	23.5	45.79	54.42	57.51	38.16	9.21	17.53	33.96	57.98	65.67	69.13
Ours	0.22	71.14	3.35	11.67	**29.98**	**48.82**	**56.62**	60.07	37.42	9.47	17.53	34.9	60.02	67.66	**70.02**
Ours + Refinement		**71.02**	**3.61**	**12.66**	29.3	48.35	56.36	**60.13**	**37.15**	**11.98**	**25.12**	**43.8**	**62.27**	**68.03**	69.96

Search Radius: 10m

experiment	uncertainty	xy error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
OrienterNet		4.06	**5.02**	15.65	39.72	69.07	91.05		12.58	11.25	21.04	40.03	75.82	88.8	92.67
Ours (32m branch)	0.26	4.11	3.87	14.91	37.47	68.29	91		12.98	11.2	20.36	39.87	74.1	88.96	92.67
Ours (128m branch)	0.13	3.95	3.77	11.36	34.64	70.96	**93.56**		8.11	13.24	25.01	46.89	81.79	92.78	**95.55**
Ours	0.2	**3.83**	4.24	**16.06**	40.55	**71.27**	92.83		8.5	13.55	26.01	49.5	83.73	93.41	95.5
Ours + Refinement		3.93	3.66	16.01	**40.61**	70.91	91.99		**8.09**	**17.27**	**34.9**	**60.33**	**86.66**	**93.77**	95.5

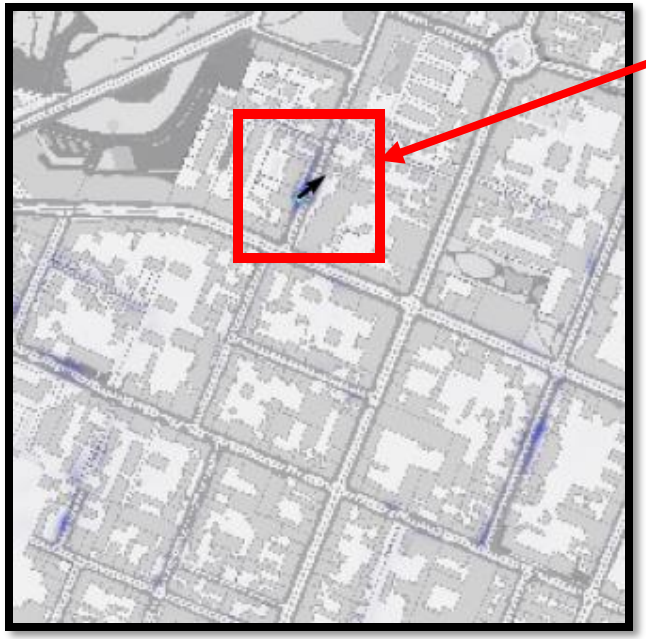
Strong improvements in orientational recall!

Hierarchical Inference

In large search areas, our fine branch requires high GPU memory.

To reduce this:

1. Match the **coarse** BEV
2. Query **fine** tile centered at argmax
3. Match the **fine** BEV.



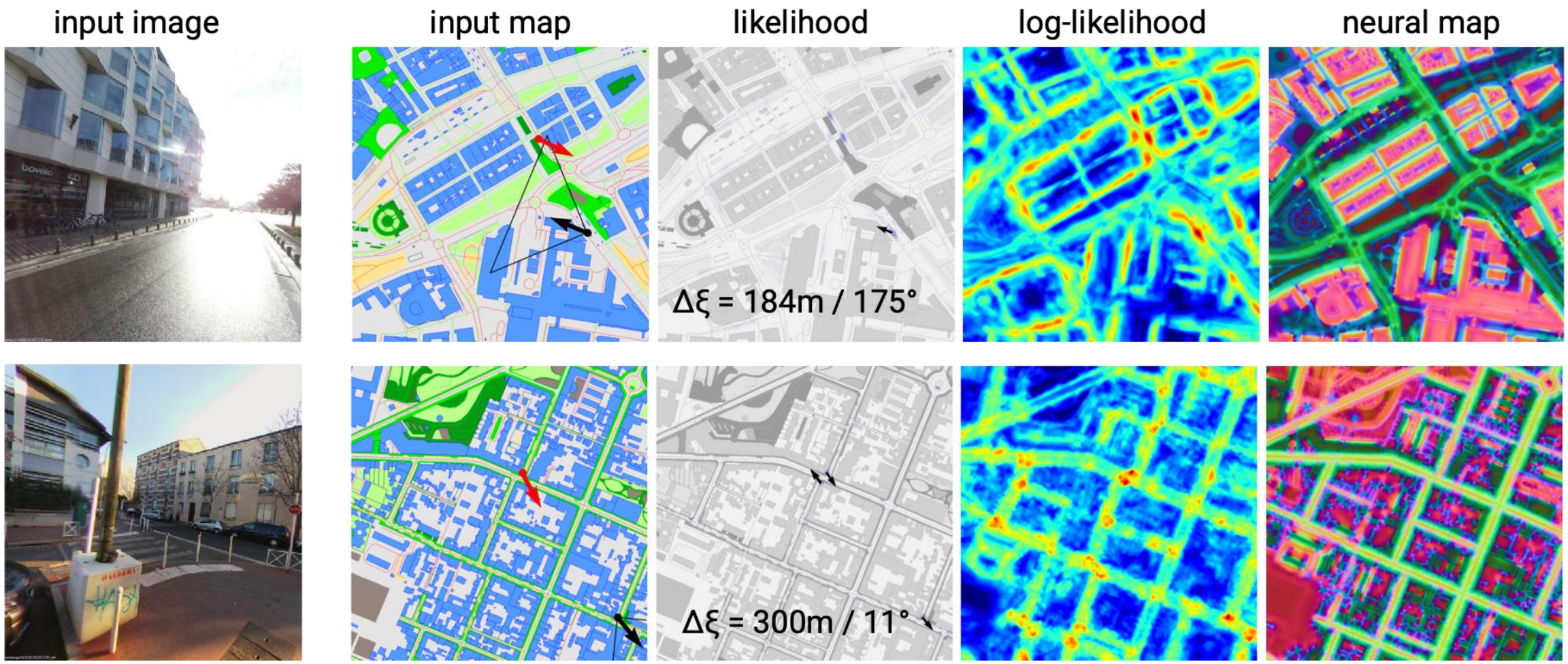
Coarse Likelihood

Reduces Memory requirement from
15GB to 3GB
when searching on map with radius
0.5km

Limitations

1. Limited information in OSM semantic maps.
2. Challenging images lacking localizable visual cues (eg. zoomed-in building wall)

Limitations



Fusing Satellite Imagery

Satellite imagery, when combined with **OSM**, results in richer neural maps.

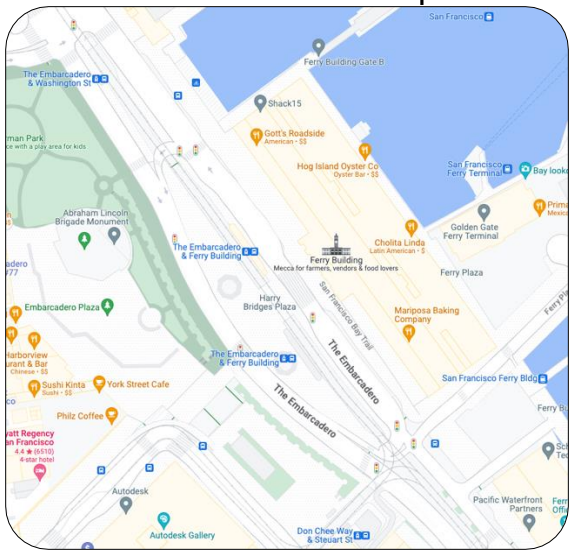
We adopt an **early fusion** approach

Satellite Imagery

Preserves **geometry,** contains more **semantics**

❌ challenging to understand

Semantic Maps



Distinct class boundaries

❌ missing information

Results with Satellite Imagery

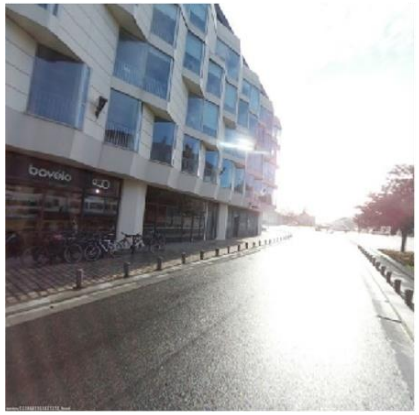
Search Radius: 256m

experiment	uncertainty	xy mean error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw mean error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
GPS	-	4.53	13.29	24.8	41.08	66.14	88.07	97.8	-	-	-	-	-	-	-
OrienterNet	-	109.58	2.83	9	22.92	34.38	38.41	41.23	52.79	6.54	12.3	23.29	42.86	50.81	54.74
Ours (32m branch)	0.33	109.5	2.15	8.11	20.83	32.86	37.57	40.87	55.6	6.12	11.25	22.19	41.81	49.92	53.43
Ours (128m branch)	0.18	78.94	2.56	7.8	23.5	45.79	54.42	57.51	38.16	9.21	17.53	33.96	57.98	65.67	69.13
Ours	0.22	71.14	3.35	11.67	29.98	48.82	56.62	60.07	37.42	9.47	17.53	34.9	60.02	67.66	70.02
Ours with Satellite (32m branch)	0.3	97.39	2.77	9.63	24.7	39.19	42.86	46.47	49.7	6.44	12.66	23.76	46.26	54.53	58.45
Ours with Satellite (128m branch)	0.18	68.52	2.98	8.84	27.32	50.13	58.97	61.9	35.08	9.79	18.42	35.06	61.17	69.34	71.9
Ours with Satellite	0.21	**64.15**	**3.66**	**13.4**	**33.07**	**53.53**	**60.65**	**64.1**	**33.2**	**10.26**	**19.36**	**36.42**	**63.58**	**70.91**	**73.42**

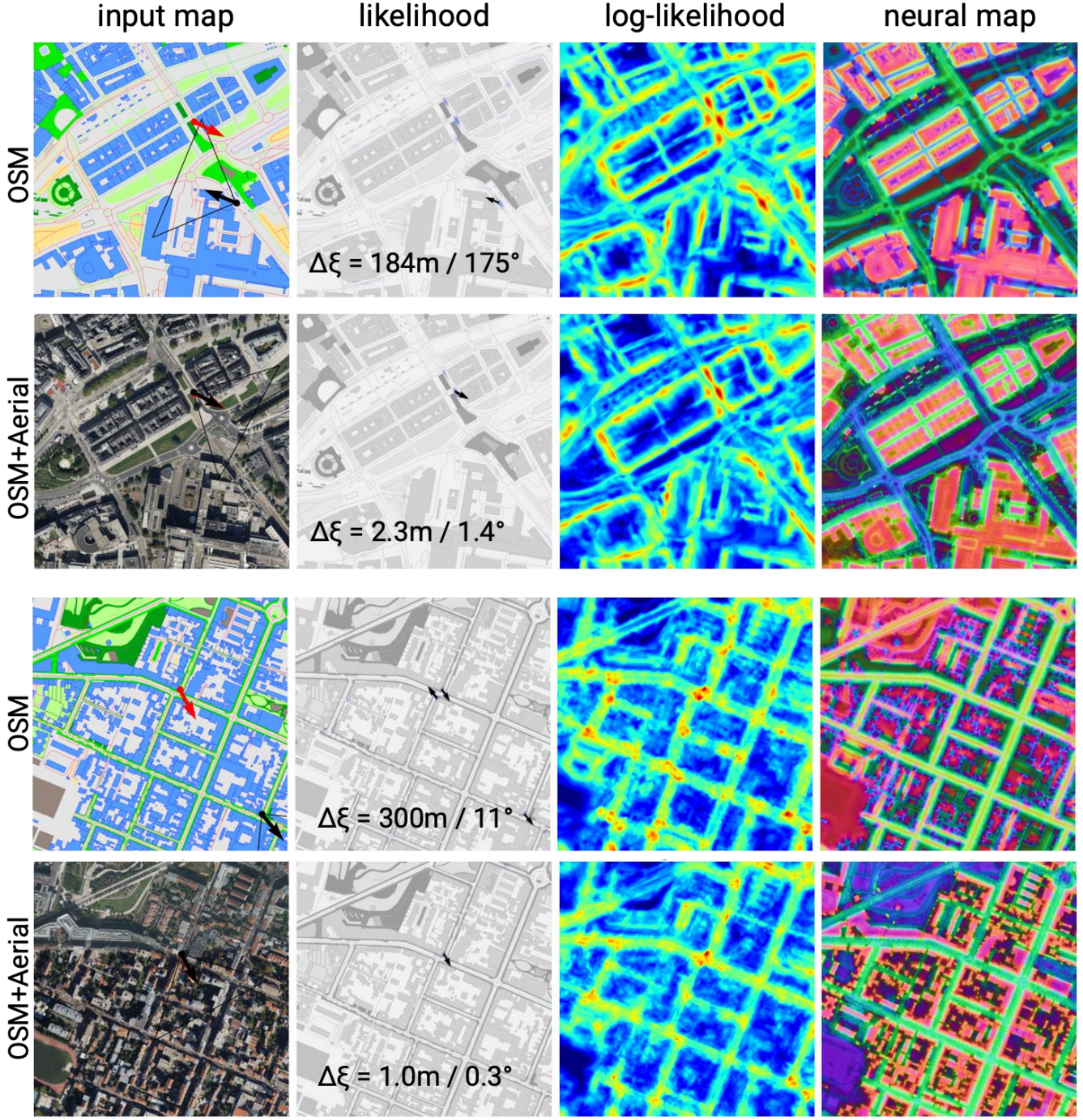
Search Radius: 10m

experiment	uncertainty	xy error	R@0.5m	R@1m	R@2m	R@5m	R@10m	R@20m	yaw error	R@0.5°	R@1°	R@2°	R@5°	R@10°	R@20°
GPS	-	4.53	13.29	24.8	41.08	66.14	88.07	97.8	-	-	-	-	-	-	-
OrienterNet		4.06	**5.02**	15.65	39.72	69.07	91.05	**100**	12.58	11.25	21.04	40.03	75.82	88.8	92.67
Ours (32m branch)	0.26	4.11	3.87	14.91	37.47	68.29	91	**100**	12.98	11.2	20.36	39.87	74.1	88.96	92.67
Ours (128m branch)	0.13	3.95	3.77	11.36	34.64	70.96	93.56	**100**	8.11	13.24	25.01	46.89	81.79	92.78	95.55
Ours	0.2	3.83	4.24	16.06	40.55	71.27	92.83	**100**	8.5	**13.55**	**26.01**	**49.5**	**83.73**	**93.41**	95.5
Ours with Satellite (32m branch)	0.24	3.96	4.19	15.02	39.46	70.59	91.68	**100**	12.39	11.04	21.04	39.4	75.04	89.01	92.62
Ours with Satellite (128m branch)	0.13	3.86	3.04	11.04	34.22	72.84	**93.93**	**100**	8.03	12.56	24.12	45.37	81.63	93.04	95.87
Ours with Satellite	0.19	**3.63**	3.72	**16.33**	**42.7**	**74.73**	93.46	**100**	**8.02**	13.03	25.33	47.51	83.36	**93.41**	**96.02**

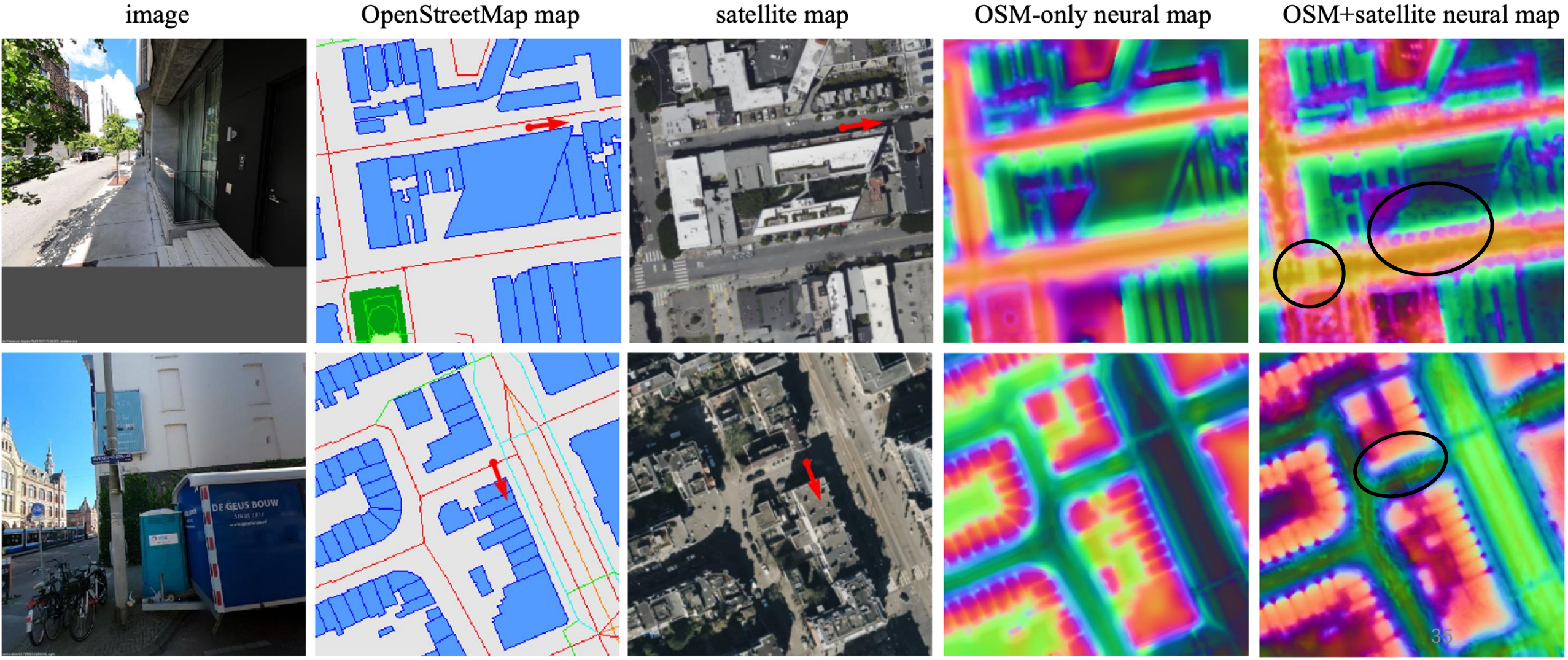
Satellite imagery resolves limitations



Road textures/markings, poles, road edges, etc. are visible in satellite maps.



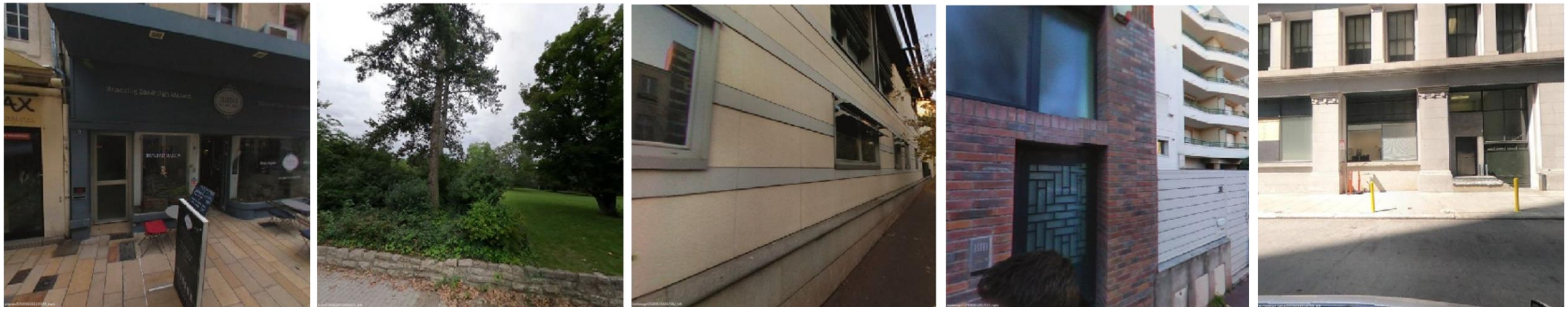
Comparing Neural Maps



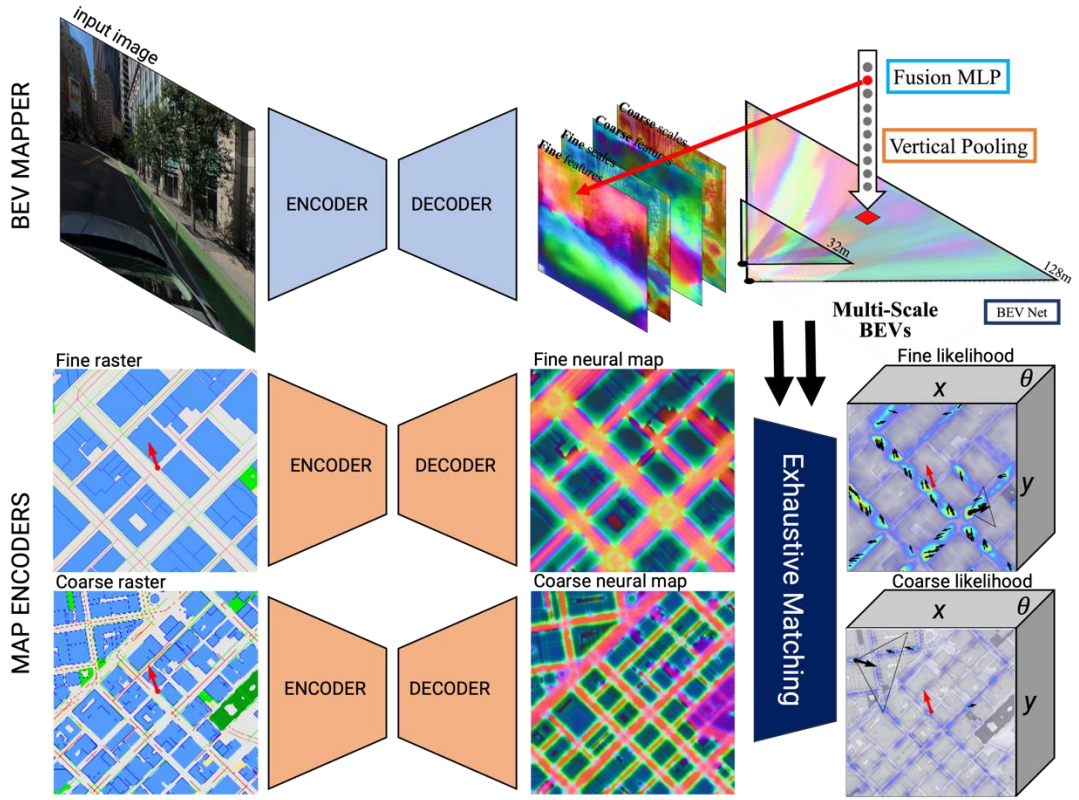
Limitations

1. Limited information in OSM semantic maps.

2. Challenging images lacking localizable visual cues (eg. zoomed-in building wall)



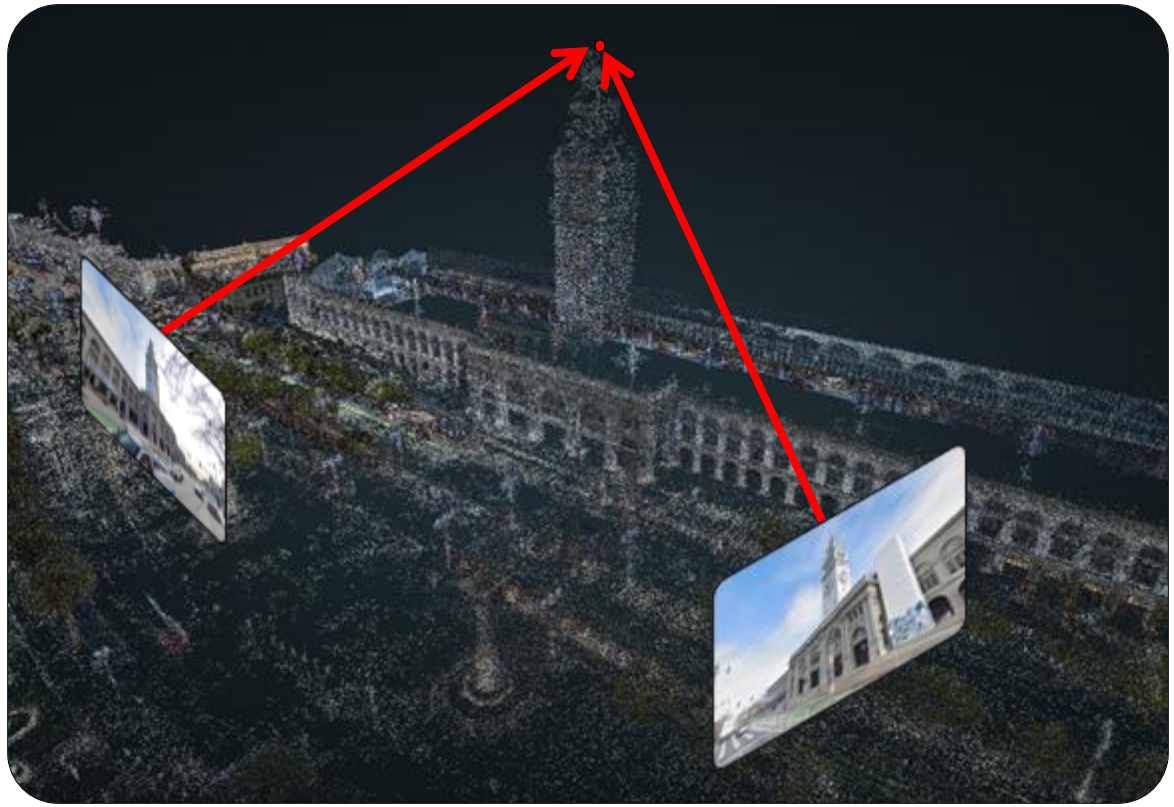
Summary



We propose a multi-scale approach that utilizes coarse and fine features from near and far to accurately estimate camera pose.

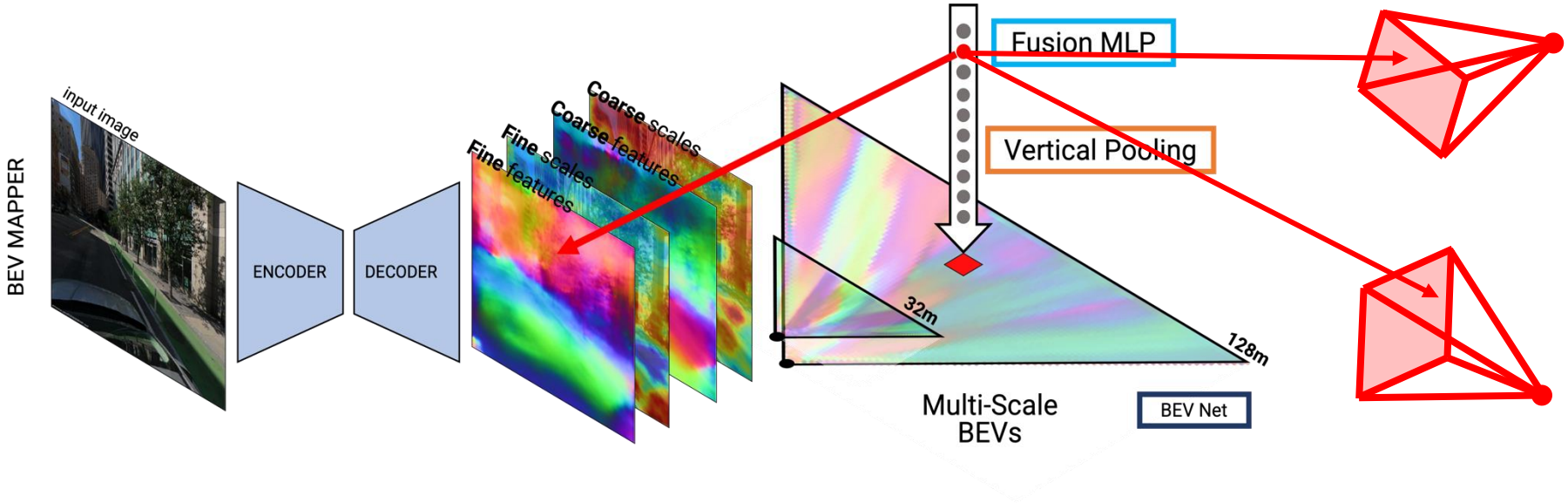
Future Directions

- Integrating into SfM



Future Directions

- Integrating into SfM
- Using multi-view constraints to improve depth and semantics



Future Directions

- Integrating into SfM
- Using multi-view constraints to improve depth and semantics
- Relaxing the assumption of known gravity direction, by jointly estimating it along with 3-DoF pose

Thank you!

Questions?

Asset credits: Paul-Edouard Sarlin, Google, and Bing Maps

Appendix

Multi-Scale Maps

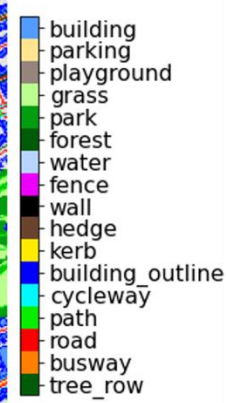
Image

Resolution: 0.5mpp
Area: 128m x 128m

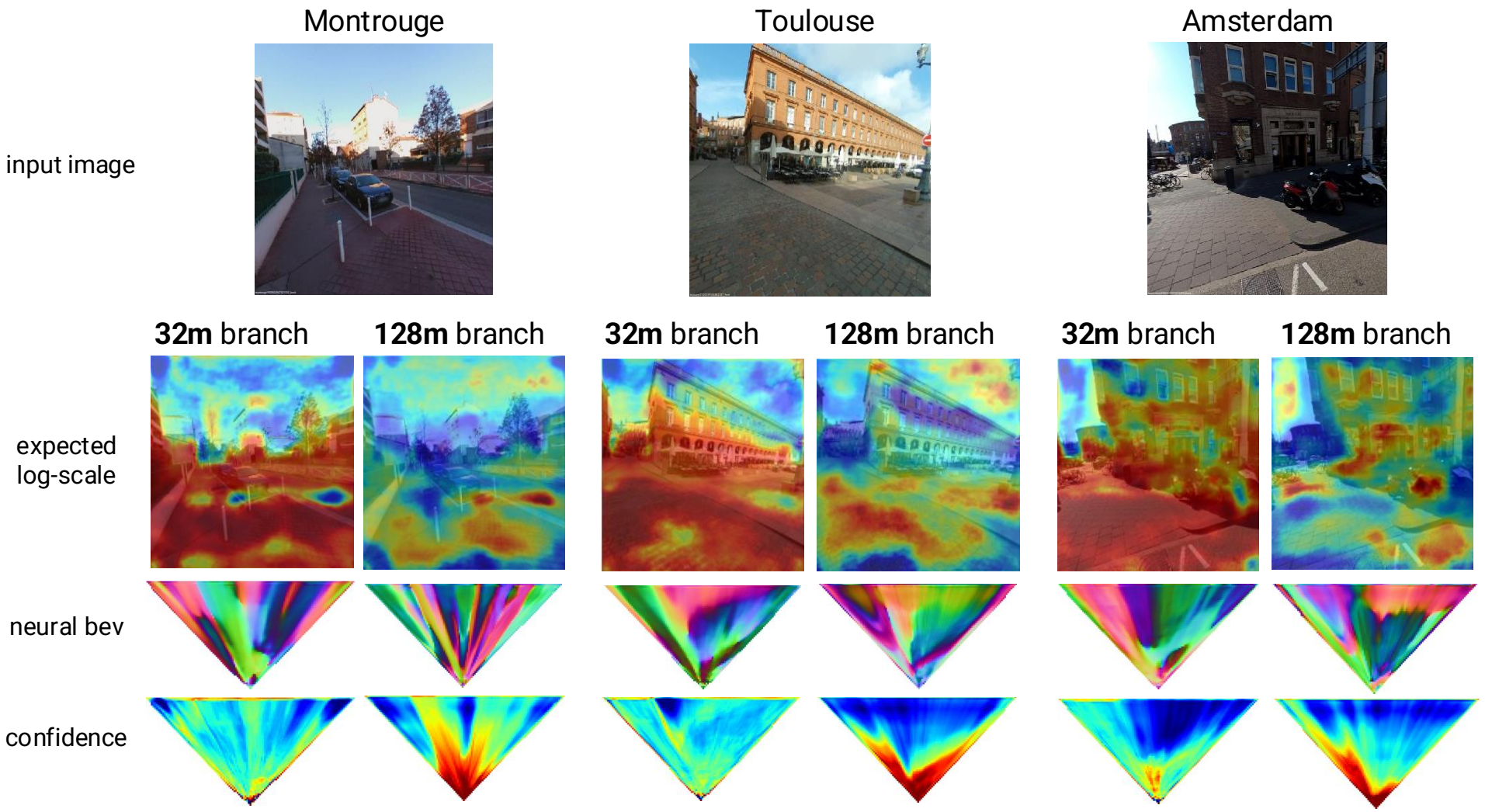
Resolution: 1mpp
Area: 256m x 256m

Resolution: 2mpp
Area: 512m x 512m

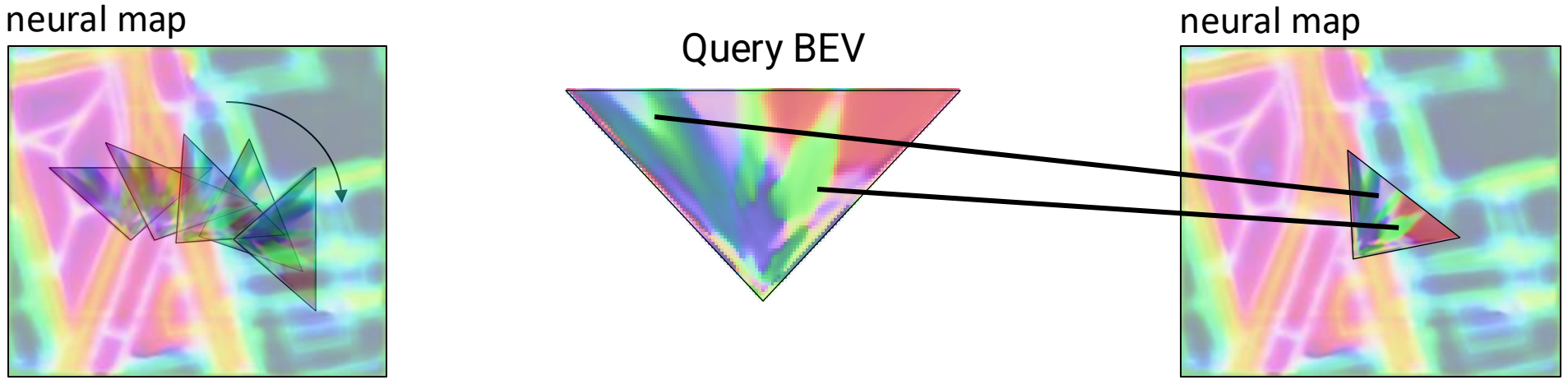
Resolution: 4mpp
Area: 1024m x 1024m



BEV Mapper


Montrouge
Toulouse
Amsterdam
input image
32m branch
128m branch
32m branch
128m branch
32m branch
128m branch
expected log-scale
neural bev
confidence

Pose Estimation: Why Exhaustive?



Exhaustive Matching

RANSAC Matching

While scoring **4M** poses

- ✅ Recall @ 5m/5° = 100%/100%
- ✅ Time: 20ms
- ✅ GPU: 2GB
- ✅ Scales well wrt. map size

While scoring **10k** poses

- ❌ Recall @ 5m/5° = 92%/60%
- ❌ Time: 70ms
- ❌ GPU: 7GB
- ❌ Scales poorly wrt. map size

Analyzing Exhaustive vs RANSAC