Part III
Learning-based Visual Localization

Eric Brachmann

Eric Brachmann

Staff Scientist

ebrach.github.io

@eric_brachmann

ebrachmann@nianticlabs.com

linkedin.com/in/eric-brachmann

NIANTIC

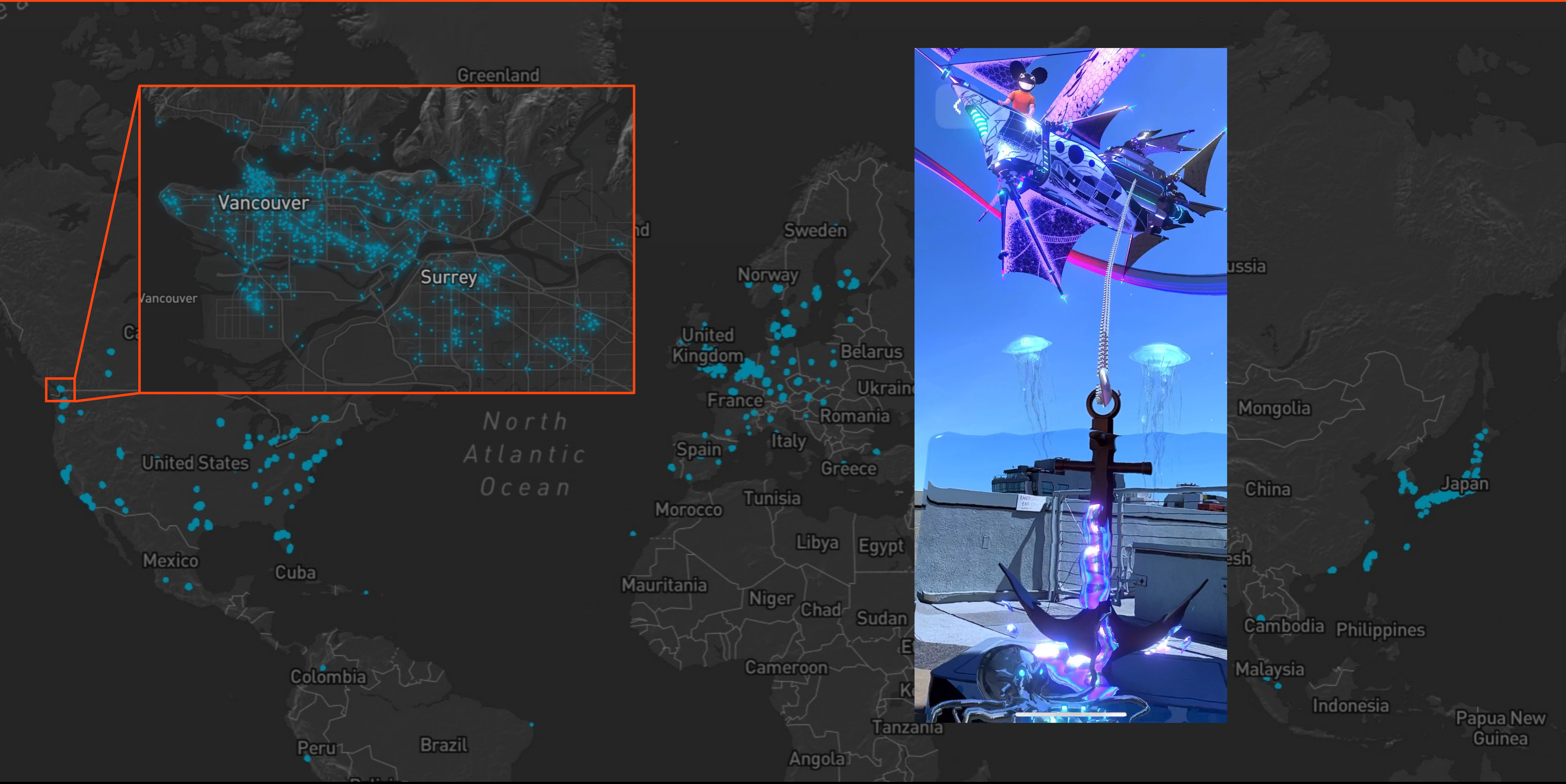
2001
Keyhole founded

2004
Keyhole acquired (becomes Google Earth)

2011
Niantic incubated at Google

2015
Niantic spins out of Google

2021
Niantic launches Lightship Platform



Greenland
Vancouver
Surrey
Vancouver
United States
Mexico
Cuba
Colombia
Peru
Brazil
North Atlantic Ocean
Sweden
Norway
United Kingdom
France
Spain
Italy
Belarus
Ukrain
Romania
Greece
Tunisia
Morocco
Libya
Egypt
Mauritania
Niger
Chad
Sudan
Cameroon
Tanzania
Angola
ussia
Mongolia
China
Japan
Cambodia
Philippines
Malaysia
Indonesia
Papua New Guinea

Part III

Regression-based

~~Learning-based~~ Visual Localization

Eric Brachmann

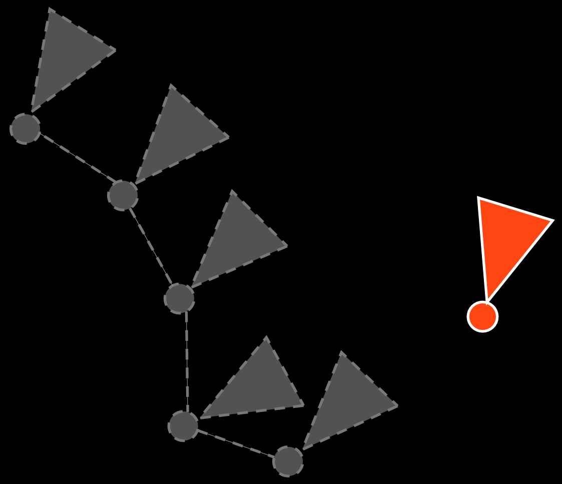
Regression

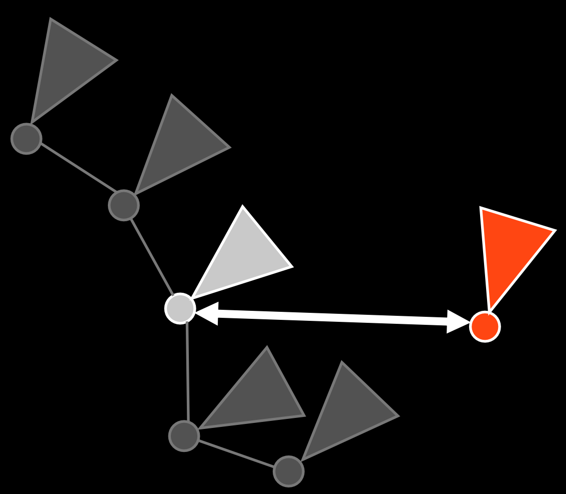
Pose Regression

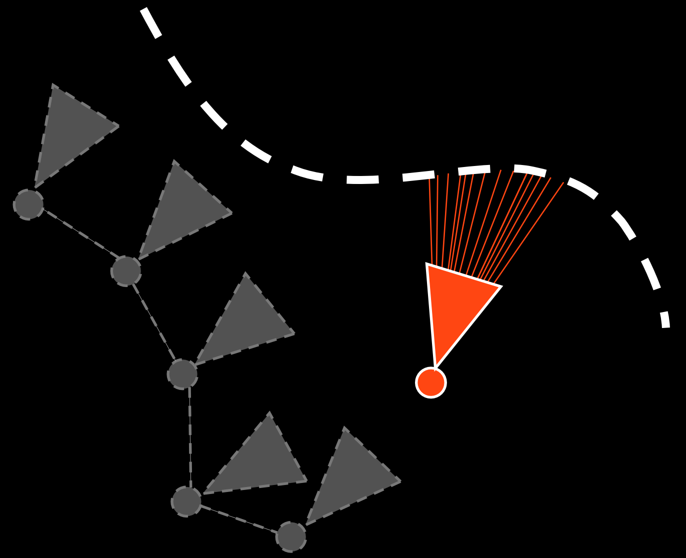
Absolute Pose Regression

Relative Pose Regression

Correspondence Regression
(aka Scene Coordinate Regression)

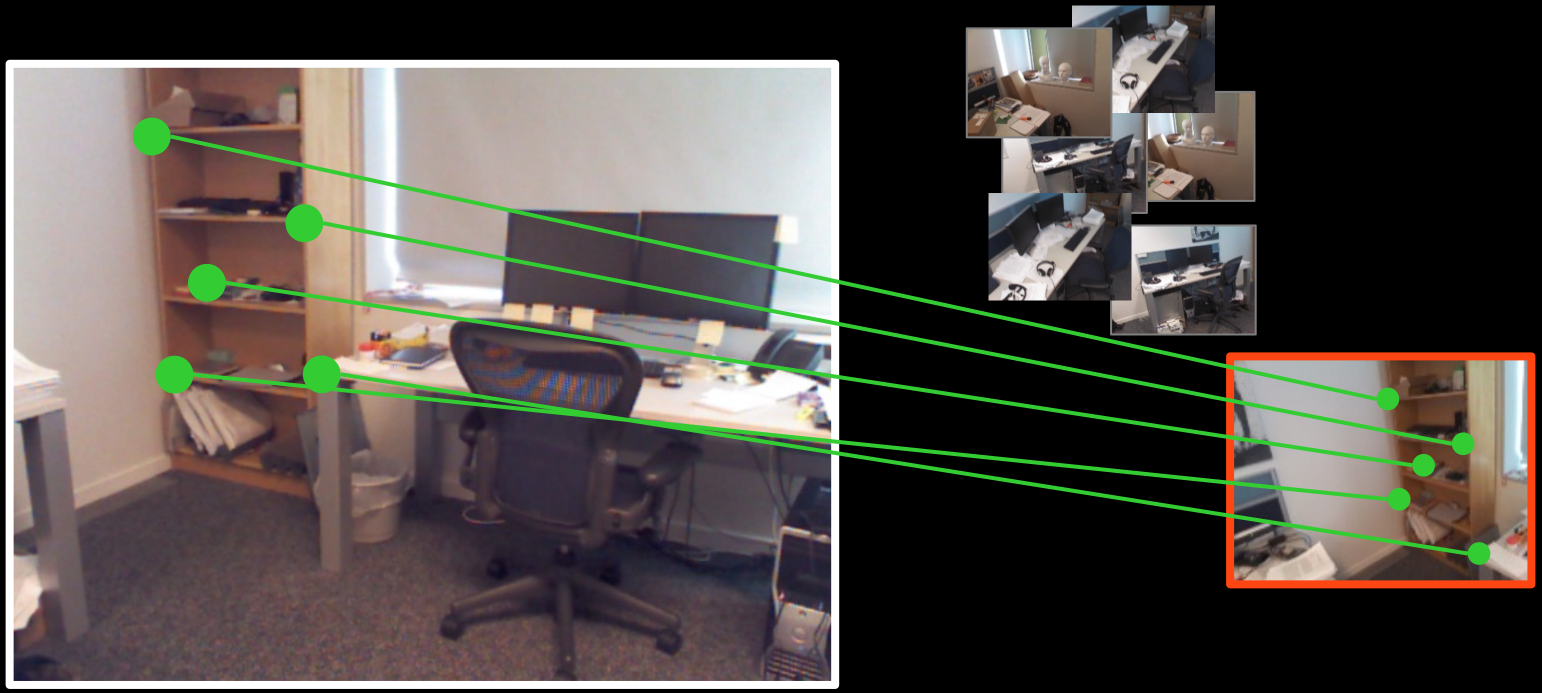


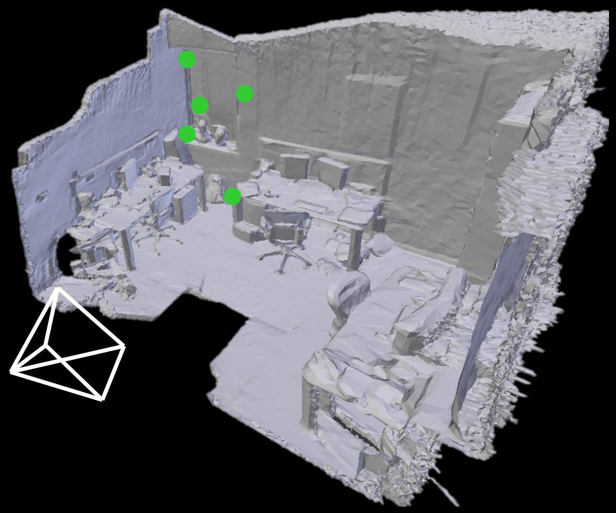




A Standard Feature Matching Pipeline

Query Image I	Image Retrieval	Feature Extraction	Feature Matching	RANSAC: Estimate Poses (PnP), Score Poses (Inliers)	6D Pose $\mathbf{h}$





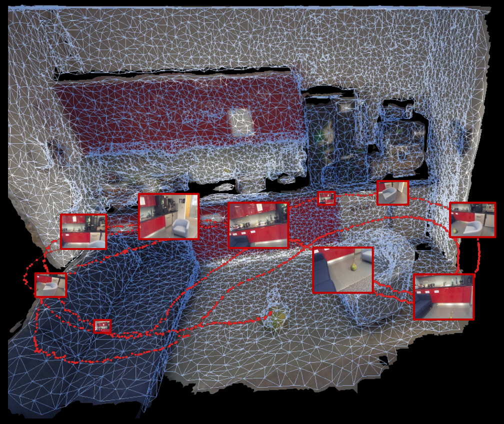
e.g. “From Coarse to Fine: Robust Hierarchical Localization at Large Scale”, Sarlin et al., CVPR’19

Query Image I	Image Retrieval	Feature Extraction	Feature Matching	RANSAC: Estimate Poses (PnP); Score Poses (Inliers)	6D Pose $\mathbf{h}$
Query Image I	Absolute Pose Regression				6D Pose $\mathbf{h}$
Query Image I	Image Retrieval	Relative Pose Regression			6D Pose $\mathbf{h}$
Query Image I	Image Retrieval	Correspondence Regression		RANSAC: Estimate Poses (PnP); Score Poses (Inliers)	6D Pose $\mathbf{h}$

Preparation
Scene-Agnostic Training

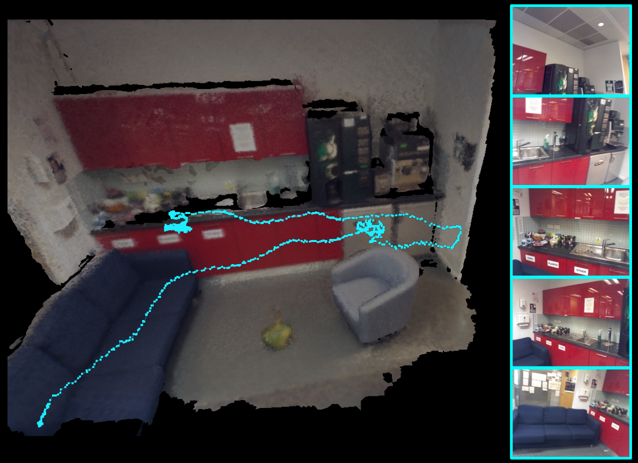
Train SuperPoint
Train SuperGlue
Train NetVLAD

Mapping
Build the Scene Representation



Obtain Posed Images
Build Retrieval Index
Triangulate Scene

Re-Localisation
Register Query Frames



NN Retrieval
Discrete Feature Matching

hLoc

[hLoc] "From Coarse to Fine: Robust Hierarchical Localization at Large Scale", Sarlin et al., CVPR'19
[SuperPoint] "SuperPoint: Self-Supervised Interest Point Detection and Description", DeTone et al., CVPR Workshops'18
[SuperGlue] "SuperGlue: Learning Feature Matching with Graph Neural Networks", Sarlin et al., CVPR'20
[NetVLAD] "NetVLAD: CNN architecture for weakly supervised place recognition", Arandjelovic et al., CVPR'16

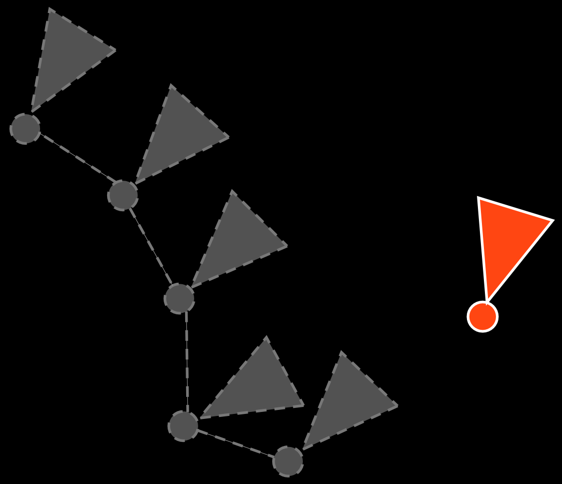
Regression

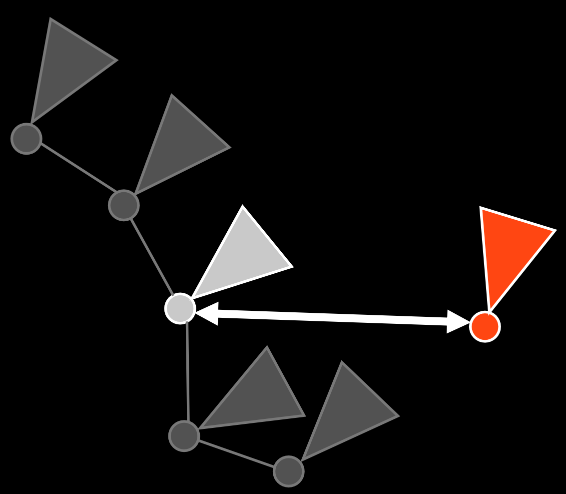
Pose Regression

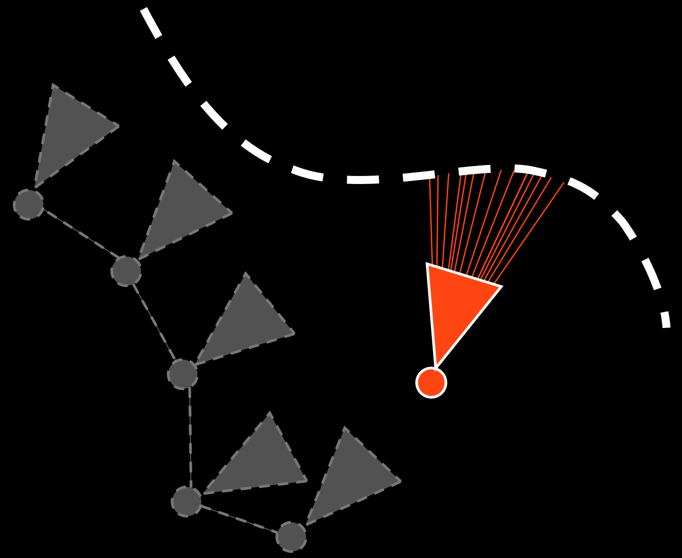
Absolute Pose Regression

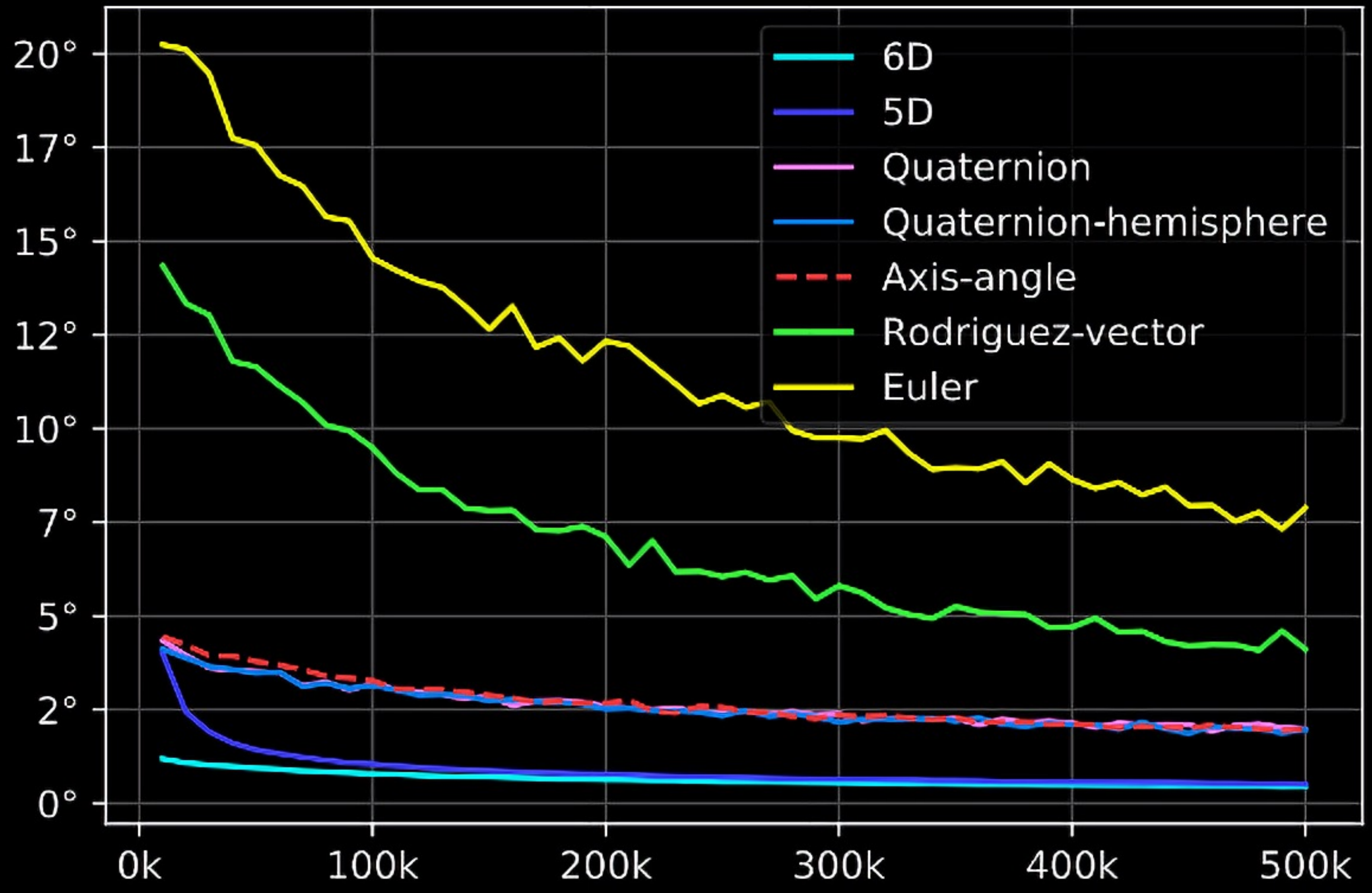
Relative Pose Regression

Correspondence Regression
(aka Scene Coordinate Regression)





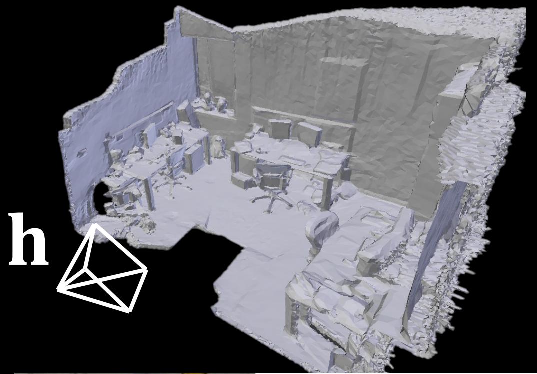


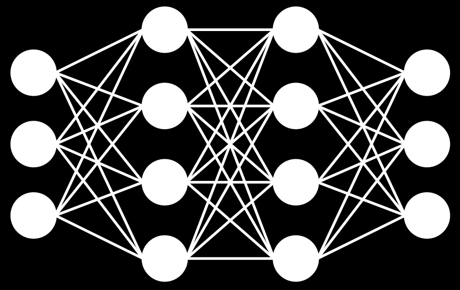


Zhou et al., "On the Continuity of Rotation Representations in Neural Networks", CVPR'19

Pose: $\mathbf{h} \in \mathrm{SE}(3) = \left\{ \begin{bmatrix} R & \mathbf{t} \\ 0 & 1 \end{bmatrix} \middle| R \in SO(3), \mathbf{t} \in \mathbb{R}^3 \right\}$

Rotation: $R \in \mathrm{SO}(3) = \{R \in \mathbb{R}^{3\times3} | RR^{\mathrm{T}} = I, \det(\mathrm{R}) = 1\}$





Rotation Matrix

$R = \begin{pmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{pmatrix}$

Enforce/Map:
$RR^{\mathrm{T}} = I, \det(\mathrm{R}) = 1$

predict 6D (e.g. in [3]) → Gram-Schmidt
predict 9D (e.g. in [4]) → orthogonal Procrustes

Unit Quaternion
e.g. [1]

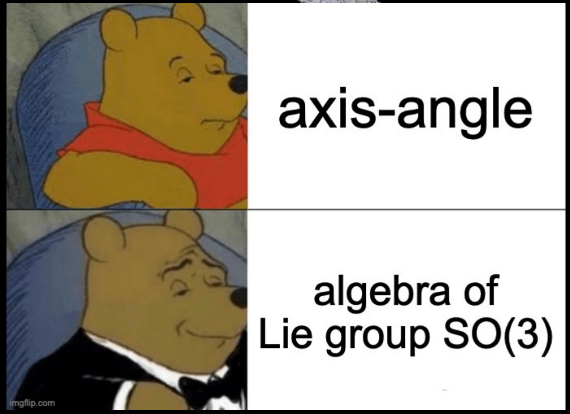
$\mathbf{q} = (c, v_1, v_2, v_3)$

Enforce/Map:
$\|\mathbf{q}\| = 1$

Axis-Angle
e.g. [2]

$\log R = \theta\hat{\mathbf{u}} = (u_1', u_2', u_3')$

Enforce/Map:
–



[1] Kendall et al., "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization", ICCV15
[2] Brachmann et al., "DSAC - Differentiable RANSAC for Camera Localization", CVPR17
[3] Zhou et al., "On the Continuity of Rotation Representations in Neural Networks", CVPR'19
[4] Chen et al., "Wide-Baseline Relative Camera Pose Estimation with Directional Learning", CVPR'21

Recommended reading:
Sola, "Quaternion kinematics for the error-state Kalman filter", 2017
Hartley et al., „Rotation Averaging", IJCV13

Rotation
- 3D normalised axis + 1D angle: 4D over-parametrisation of 3D axis-angle
  - Zhou et al., "On the Continuity of Rotation Representations in Neural Networks", CVPR19
- log unit quaternion: equivalent to axis-angle up to a scale factor of 2
  - Brahmbhatt et al., "Geometry-Aware Learning of Maps for Camera Localization", CVPR18
- Euler angles
  - Zhou et al., "On the Continuity of Rotation Representations in Neural Networks", CVPR19
- discretized Euler angles: classification rather than regression
  - Cai et al., "Extreme Rotation Estimation using Dense Correlation Volumes", CVPR21

Translation:
- 3D normalised translation vector + 1D scale: 4D over-parametrisation of 3D translation
  - Arnold et al., "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22
- 2D discretized translation direction + 1D scale: classification rather than regression
  - Arnold et al., "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22

Pose:
- Three hallucinated 3D-3D correspondences (18D): get full 6D pose via Kabsch algorithm
  - Arnold et al., "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22

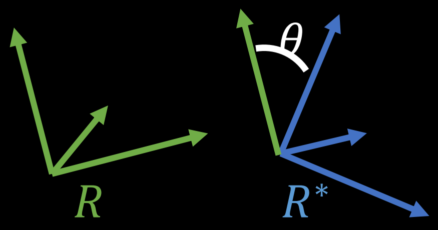
Extensive experimental analysis in "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", Arnold et al., ECCV22
Predicting rotation matrix (see previous slide) and vanilla translation vector wins.

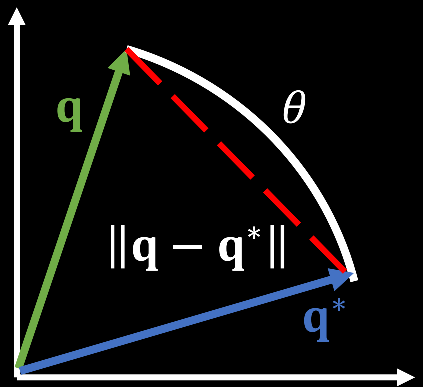
Translation error: $\ell(\mathbf{t}, \mathbf{t}^*) = \|\mathbf{t} - \mathbf{t}^*\|$

How to measure rotation error?

- Angular Distance [1]: $\theta(R, R^*) = \|\log(R^* R^{\mathrm{T}})\|$
- Quaternion Distance [2]: $\|\mathbf{q} - \mathbf{q}^*\|$
- Chordal Distance [3]: $\|\mathbf{R} - \mathbf{R}^*\|_{\mathrm{F}}$
- Angle-Axis Distance [4]: $\|\log R - \log R^*\|$

$\|\log R - \log R^*\| \neq$
$\|\log TR - \log TR^*\|$ [5]





[1] Brachmann et al., “DSAC - Differentiable RANSAC for Camera Localization”, CVPR17
[2] Kendall et al., “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization”, ICCV15
[3] Zhou et al., “On the Continuity of Rotation Representations in Neural Networks”, CVPR19
[4] Brahmbhatt et al., “Geometry-Aware Learning of Maps for Camera Localization”, CVPR18
[5] Hartley et al., „Rotation Averaging“, IJCV13

How to combine rotation error and translation error?

Weighted [1]:

$$\ell_\beta(\mathbf{h}, \mathbf{h}^*) = \ell(\mathbf{t}, \mathbf{t}^*) + \beta \ell(R, R^*)$$

Adaptive weighted [2]:

$$\ell_{\sigma^2}(\mathbf{h}, \mathbf{h}^*) = \ell(\mathbf{t}, \mathbf{t}^*) \exp(-s_t) + s_t + \ell(R, R^*) \exp(-s_R) + s_R$$

Reprojection Error [2]:

$$\ell_\pi(\mathbf{h}, \mathbf{h}^*) = \sum_{\mathbf{v} \in \mathcal{M}} \|\pi(\mathbf{h}, \mathbf{v}) - \pi(\mathbf{h}^*, \mathbf{v})\|$$

[1] Kendall et al., “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization”, ICCV15
[2] Kendall and Cipolla, “Geometric Loss Functions for Camera Pose Regression with Deep Learning”, CVPR17

Reprojection Error [2]: $\ell_\pi(\mathbf{h}, \mathbf{h}^*) = \sum_{\mathbf{v} \in \mathcal{M}} \|\pi(\mathbf{h}, \mathbf{v}) - \pi(\mathbf{h}^*, \mathbf{v})\|$

Dense Correspondence Reprojection Error (DCRE) [3]

Virtual Correspondence Reprojection Error (VCRE) [4]

- Needs depth or 3D scene model
- Mimics augmenting the scene

- No depth / 3D scene model needed
- Mimics placing virtual content

[1] Kendall et al., “PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization”, ICCV15
[2] Kendall and Cipolla, “Geometric Loss Functions for Camera Pose Regression with Deep Learning”, CVPR17
[3] Wald et al., “Beyond controlled environments: 3D camera re-localization in changing indoor scenes”, ECCV20
[4] Arnold et al., “Map-free Visual Relocalization: Metric Pose Relative to a Single Image”, ECCV22

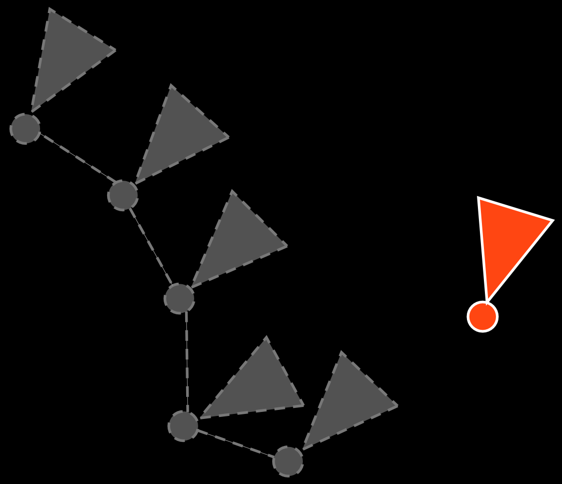
Regression

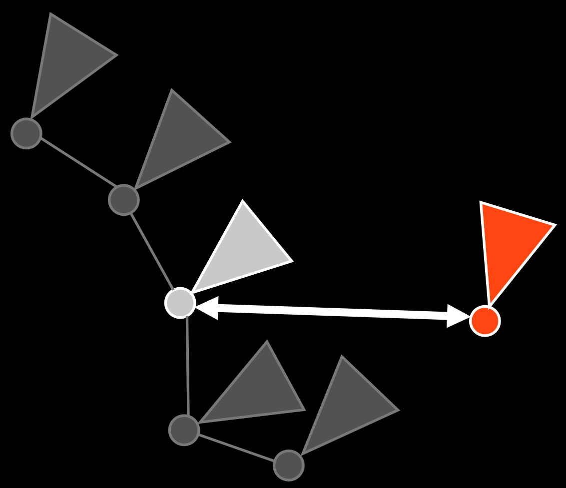
Pose Regression

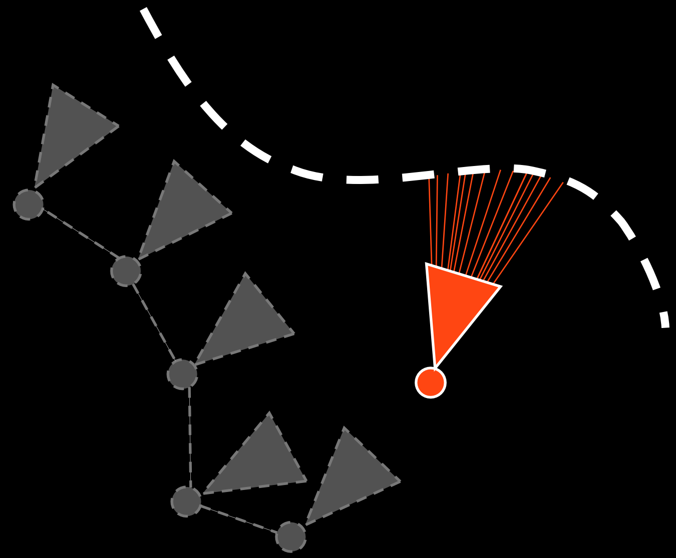
Absolute Pose Regression

Relative Pose Regression

Correspondence Regression (aka Scene Coordinate Regression)



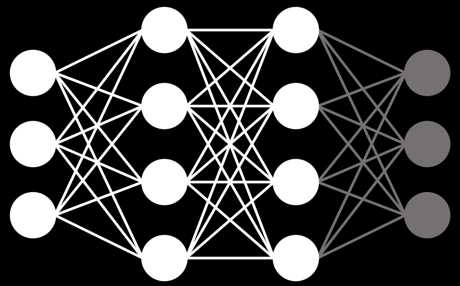




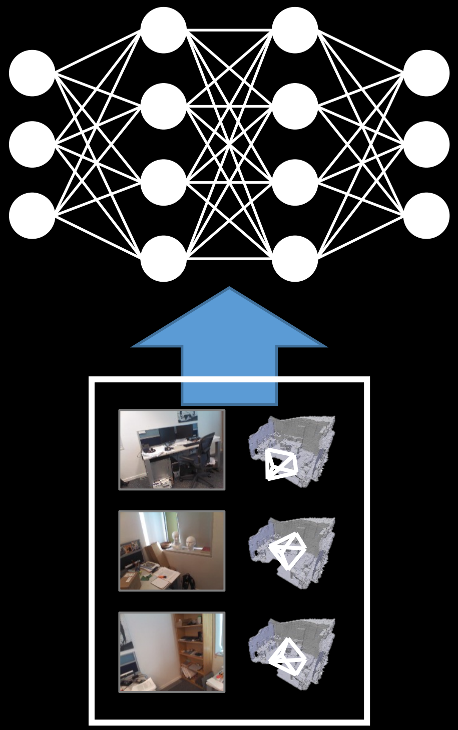
Preparation	Mapping	Re-Localisation
Scene-Agnostic Training	Scene-Specific Training	Register Query Frames

PoseNet

Pre-Train Backbone



Obtain Posed Images
Train Pose Regression

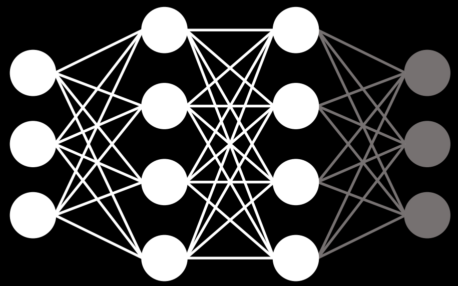


Training Data

[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17

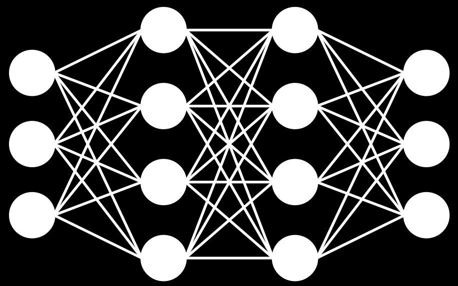
Preparation	**Mapping**	**Re-Localisation**
Scene-Agnostic Training	**Scene-Specific Training**	Register Query Frames

Pre-Train Backbone

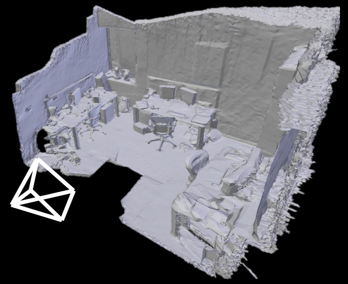


Obtain Posed Images
Train Pose Regression

Input:
Image I

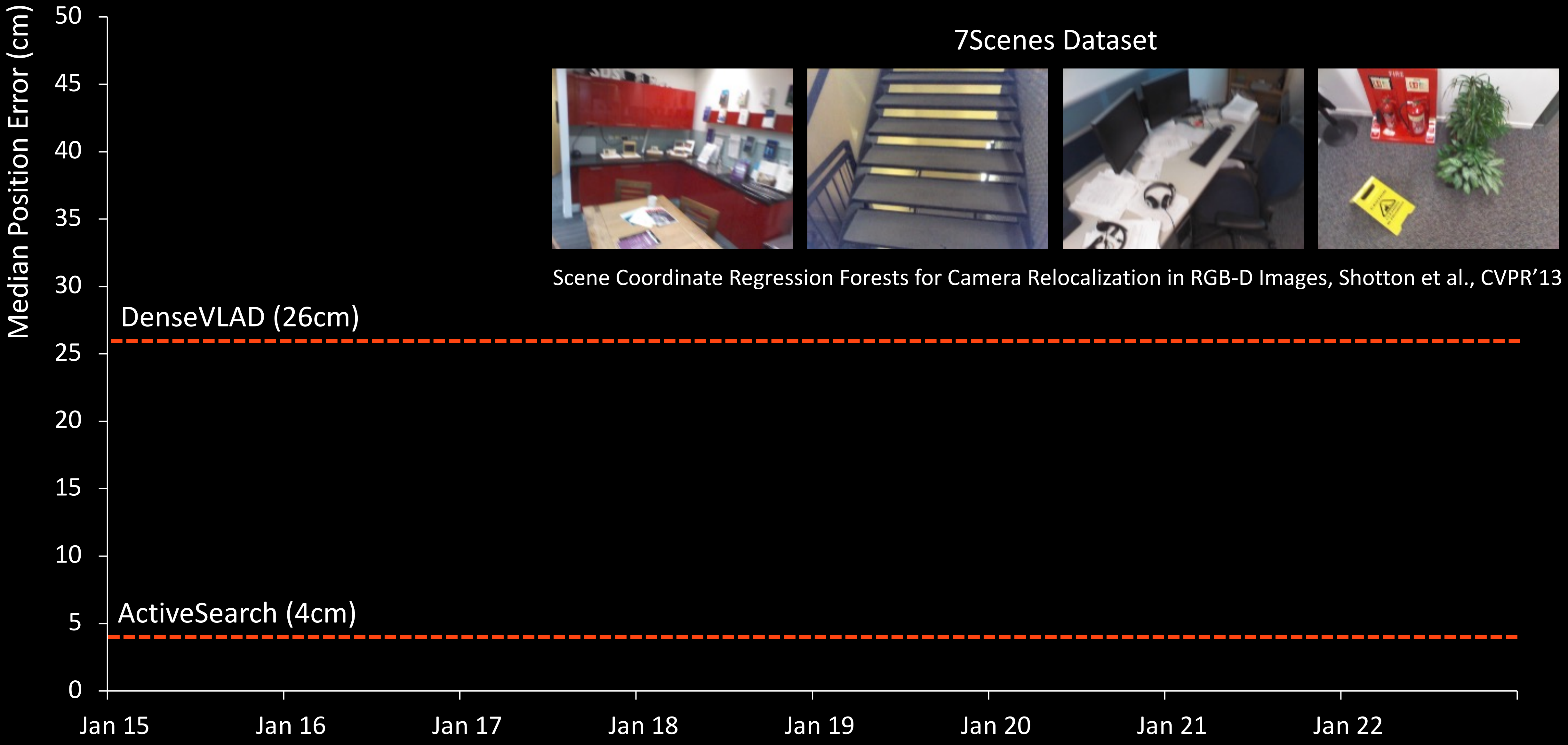


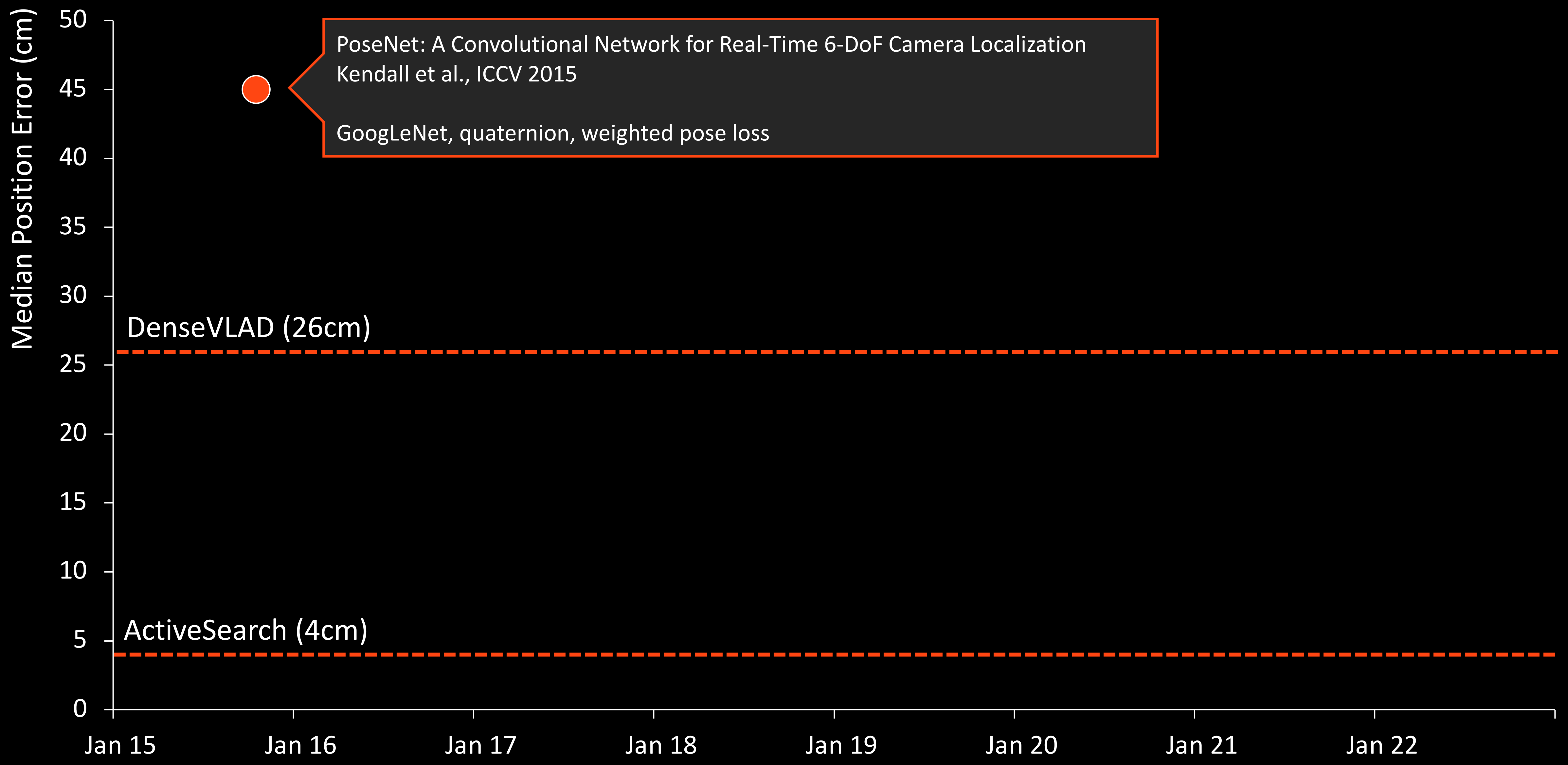
Forward Pass



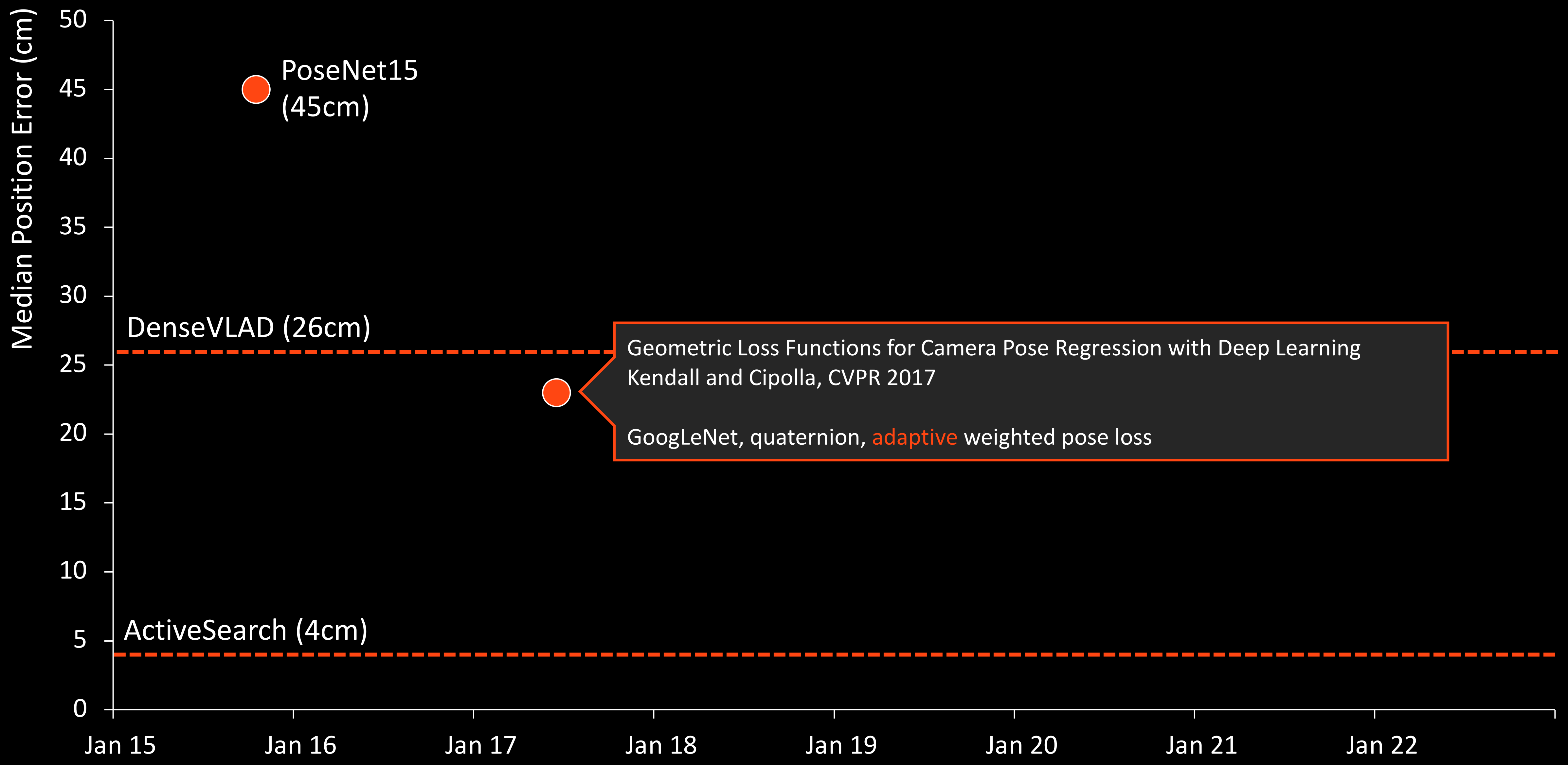
Output:
Pose $\hat{\mathbf{h}}$

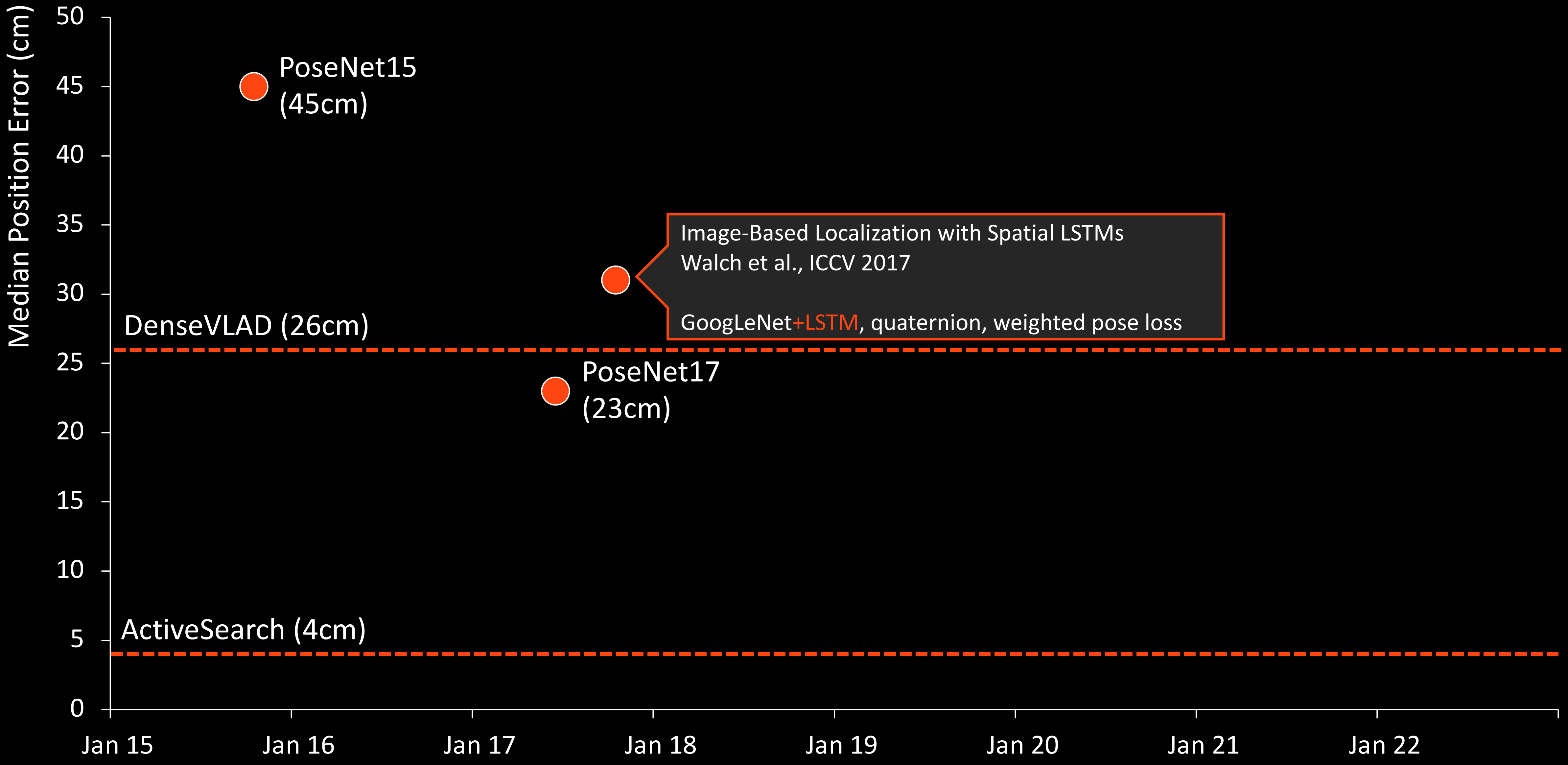
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17

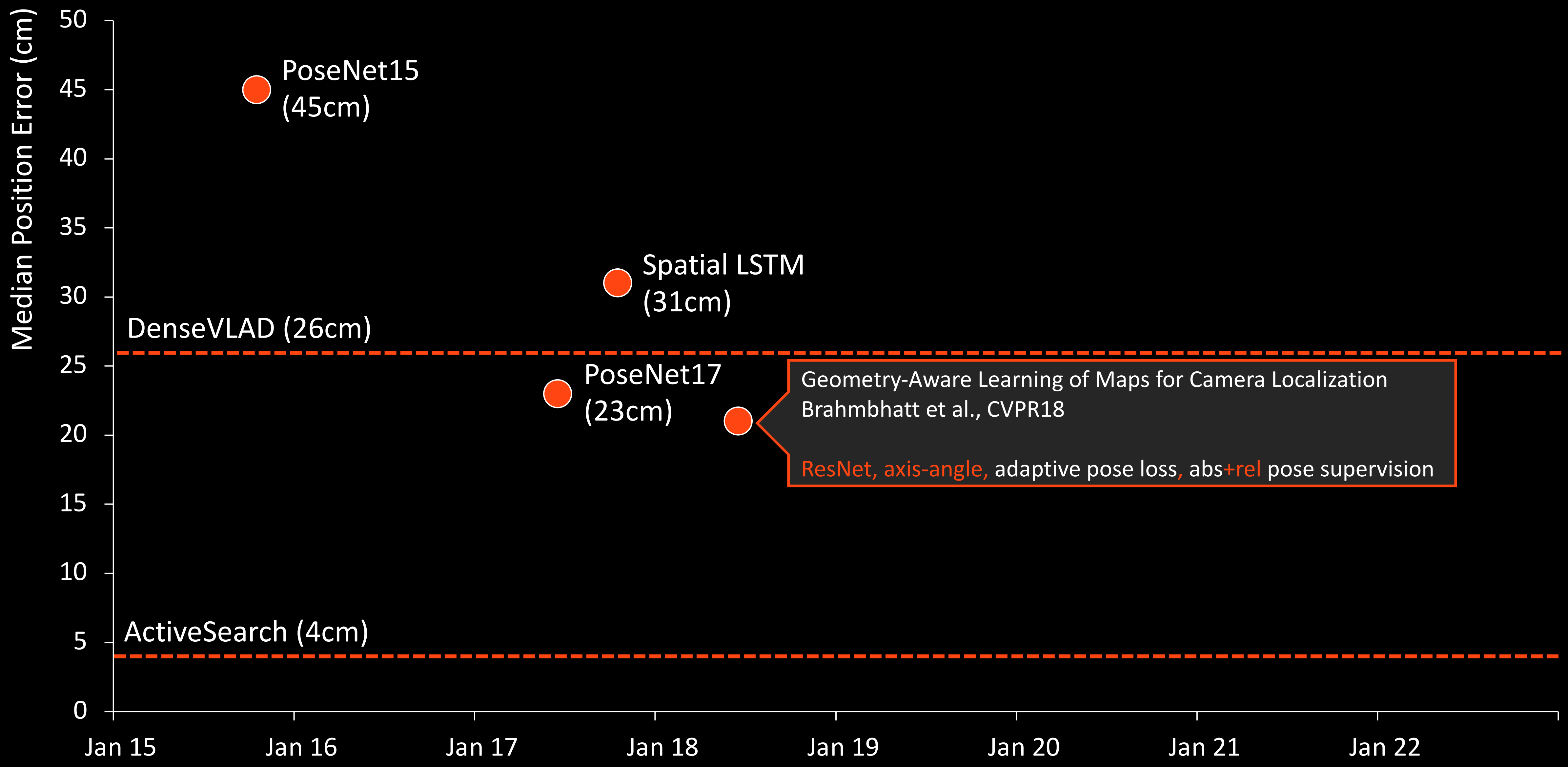

Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
7Scenes Dataset
Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR’13
DenseVLAD (26cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22



Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Localization
Kendall et al., ICCV 2015
GoogLeNet, quaternion, weighted pose loss
DenseVLAD (26cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

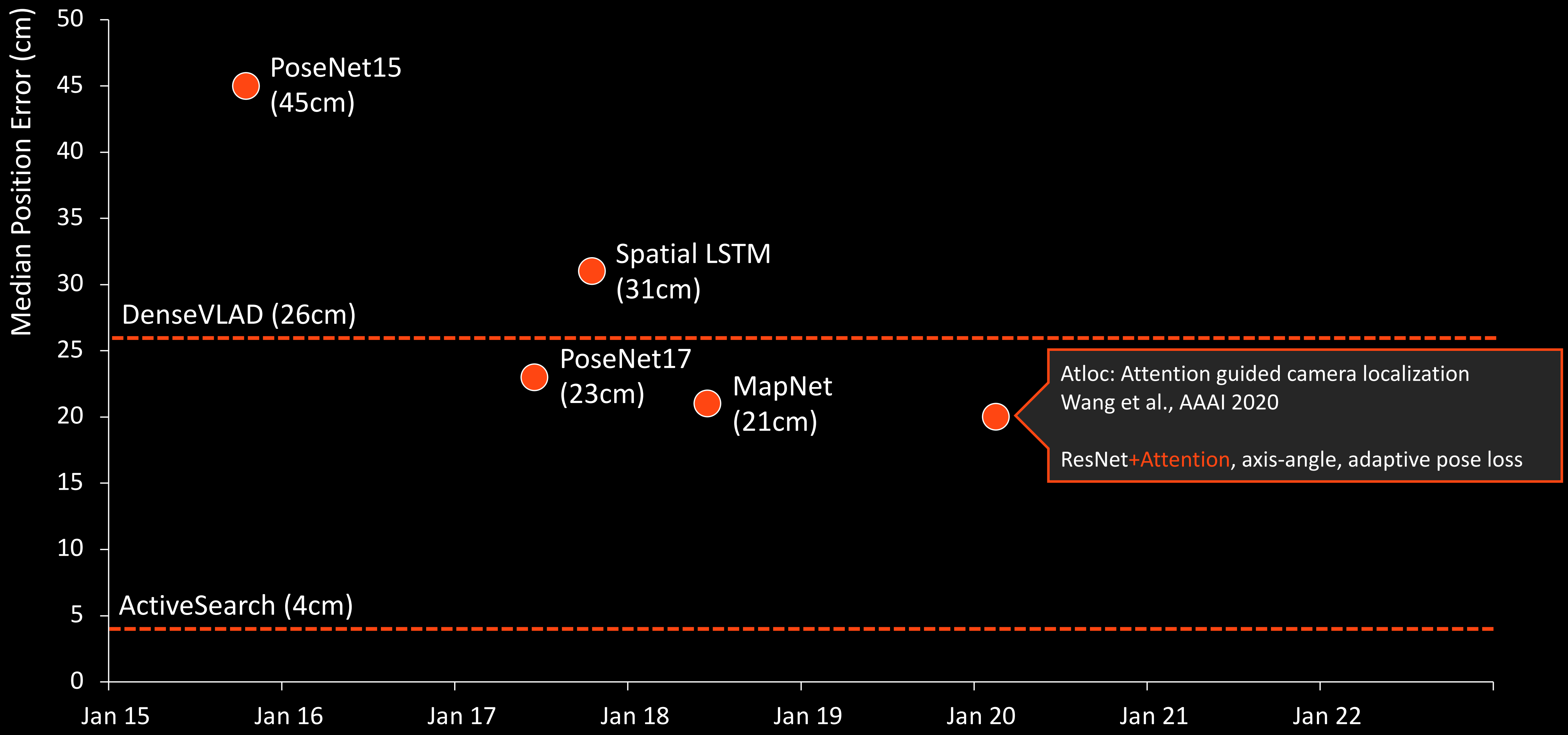

Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
DenseVLAD (26cm)
ActiveSearch (4cm)
Geometric Loss Functions for Camera Pose Regression with Deep Learning
Kendall and Cipolla, CVPR 2017
GoogLeNet, quaternion, adaptive weighted pose loss
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

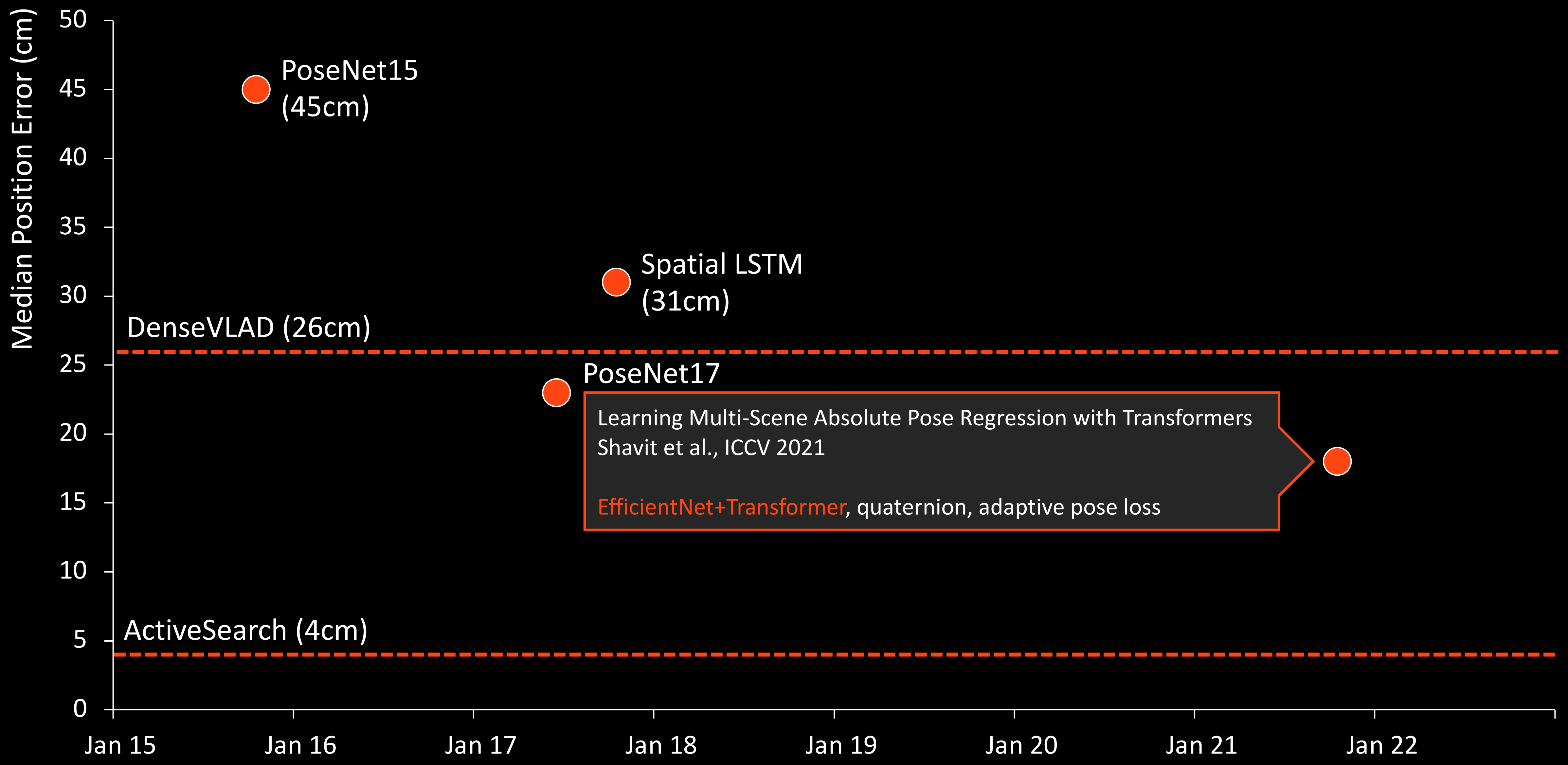

Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Image-Based Localization with Spatial LSTMs
Walch et al., ICCV 2017
GoogLeNet+LSTM, quaternion, weighted pose loss
DenseVLAD (26cm)
PoseNet17
(23cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

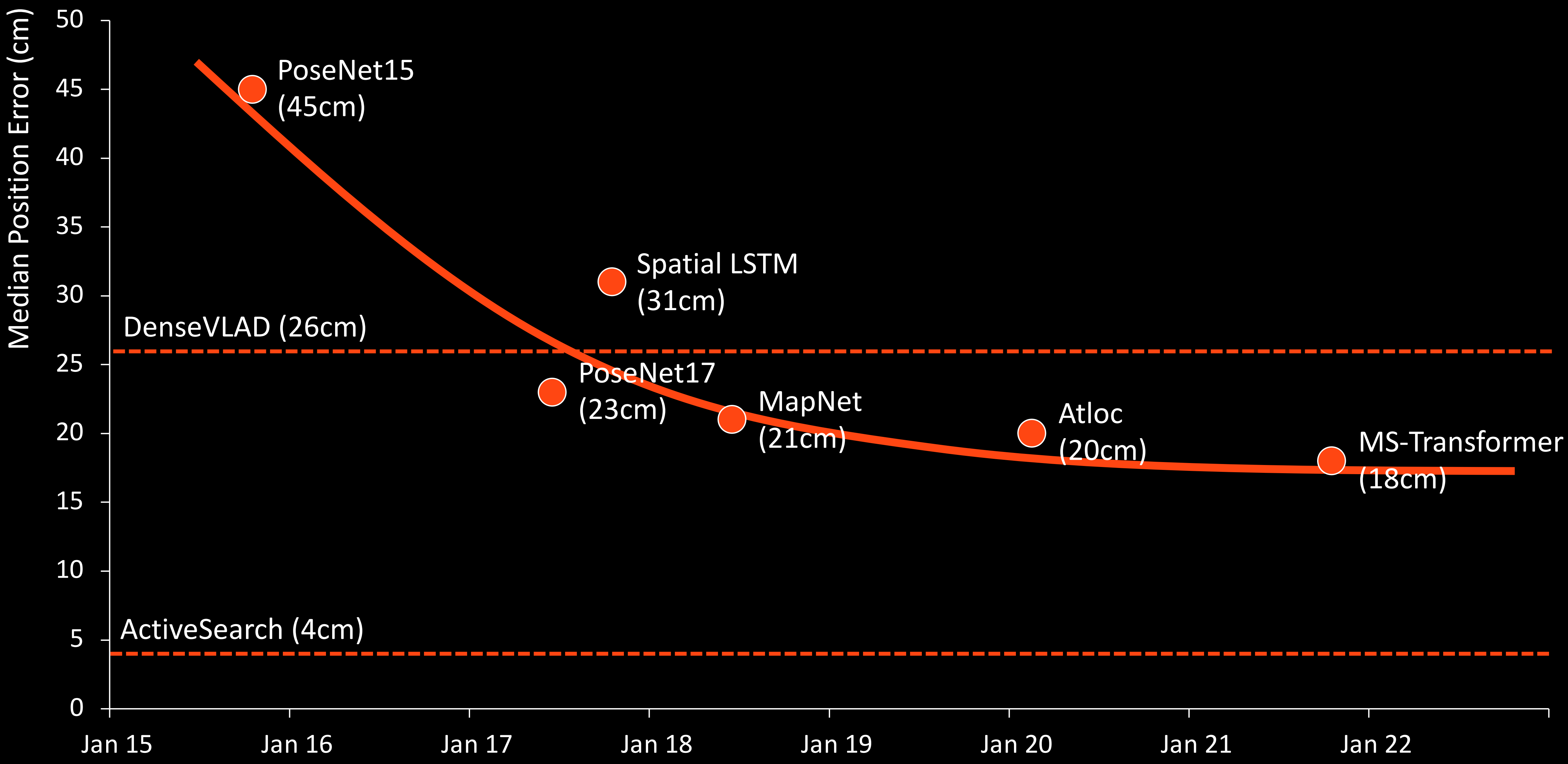


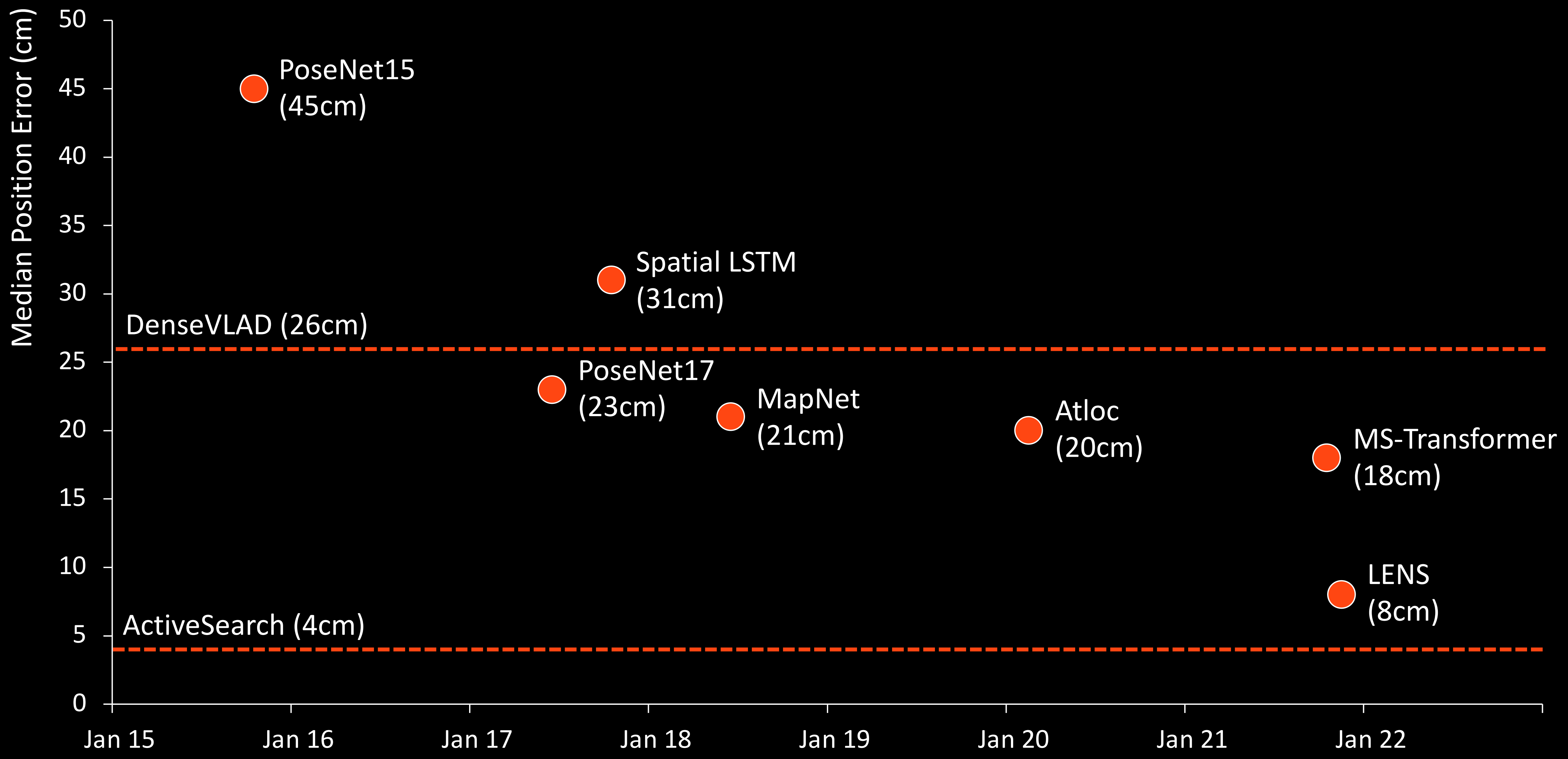
Geometry-Aware Learning of Maps for Camera Localization
Brahmbhatt et al., CVPR18

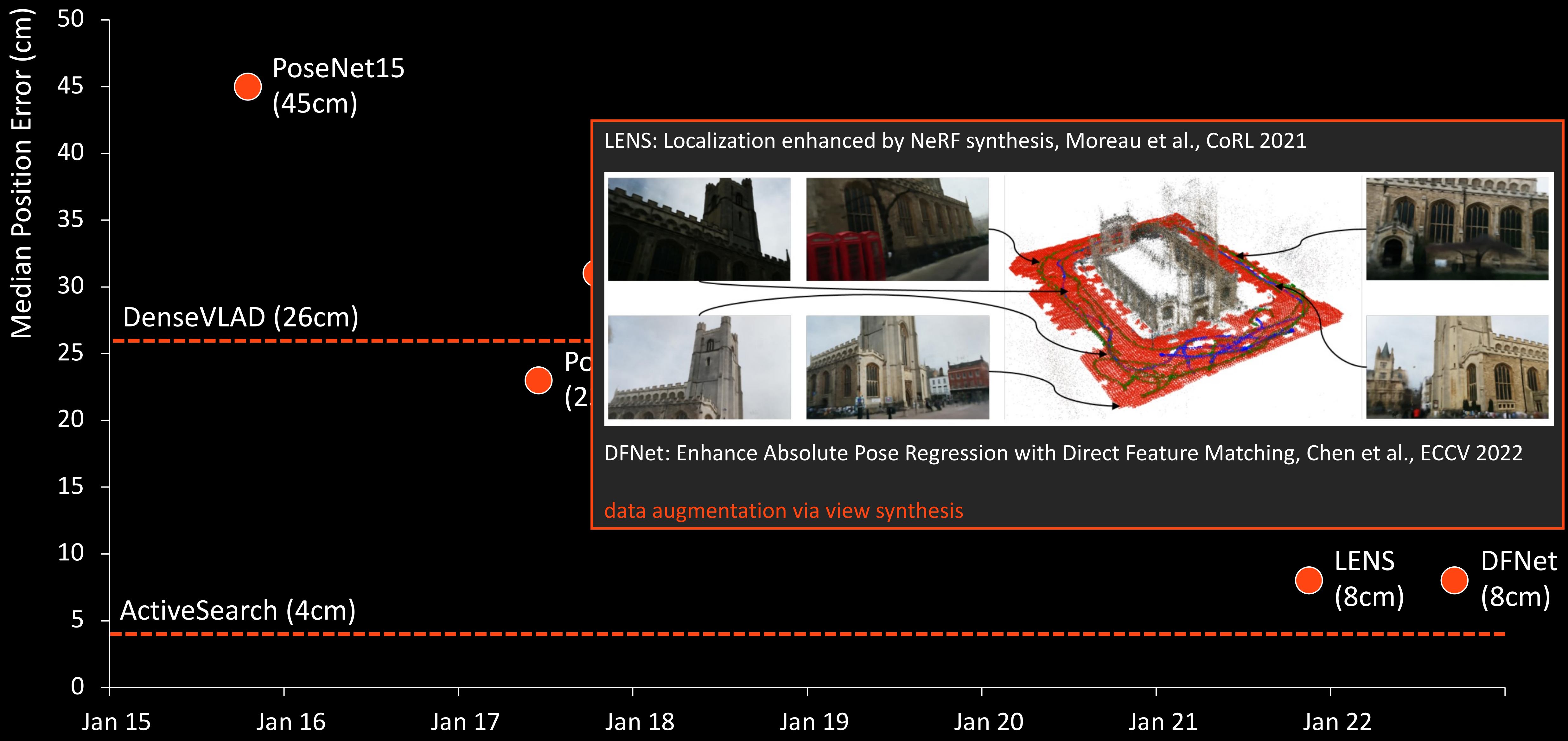
ResNet, axis-angle, adaptive pose loss, abs+rel pose supervision


Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Spatial LSTM
(31cm)
DenseVLAD (26cm)
PoseNet17
(23cm)
MapNet
(21cm)
Atloc: Attention guided camera localization
Wang et al., AAAI 2020
ResNet+Attention, axis-angle, adaptive pose loss
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22


Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Spatial LSTM
(31cm)
DenseVLAD (26cm)
PoseNet17
Learning Multi-Scene Absolute Pose Regression with Transformers
Shavit et al., ICCV 2021
EfficientNet+Transformer, quaternion, adaptive pose loss
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22


Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Spatial LSTM
(31cm)
DenseVLAD (26cm)
PoseNet17
(23cm)
MapNet
(21cm)
Atloc
(20cm)
MS-Transformer
(18cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

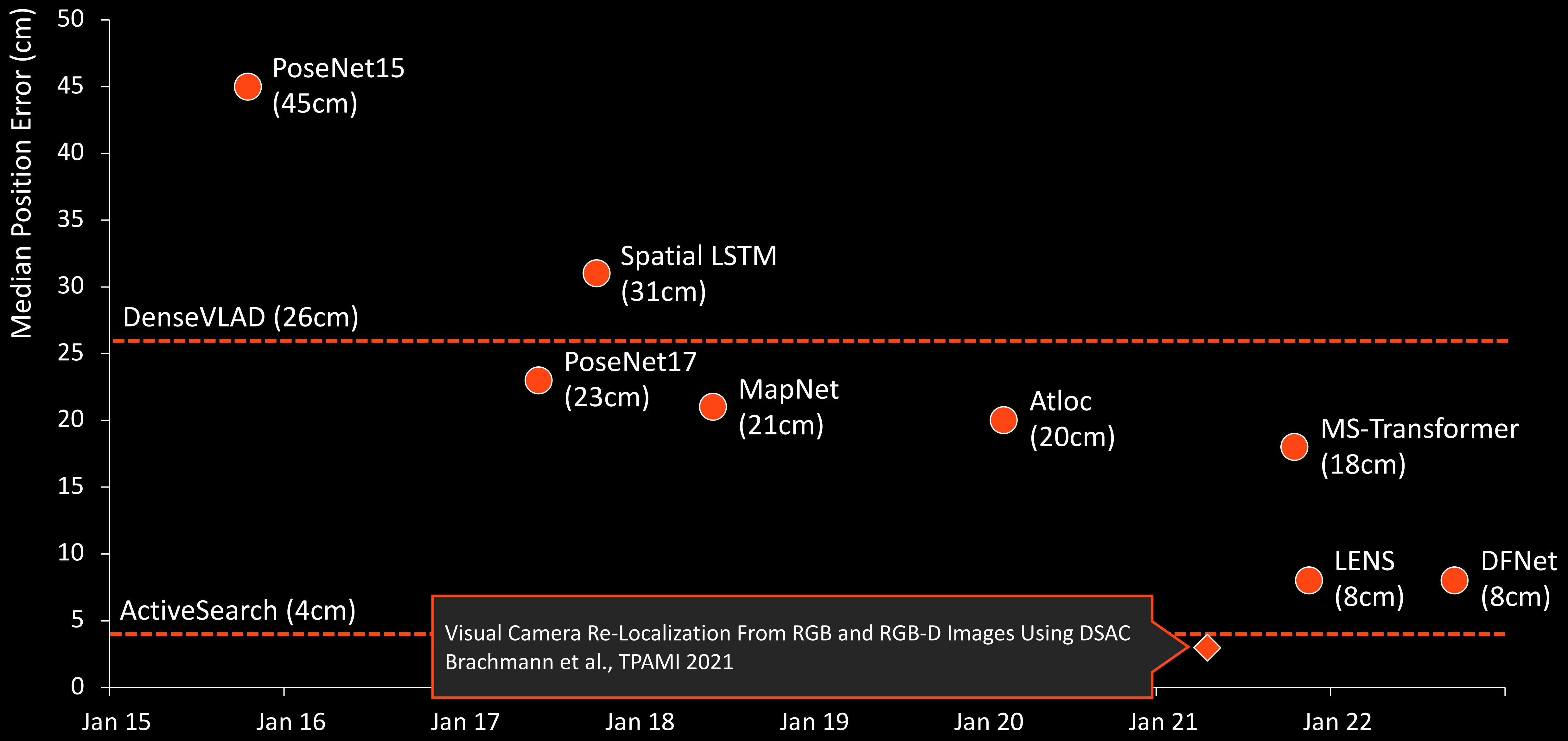

Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Spatial LSTM
(31cm)
DenseVLAD (26cm)
PoseNet17
(23cm)
MapNet
(21cm)
Atloc
(20cm)
MS-Transformer
(18cm)
LENS
(8cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22



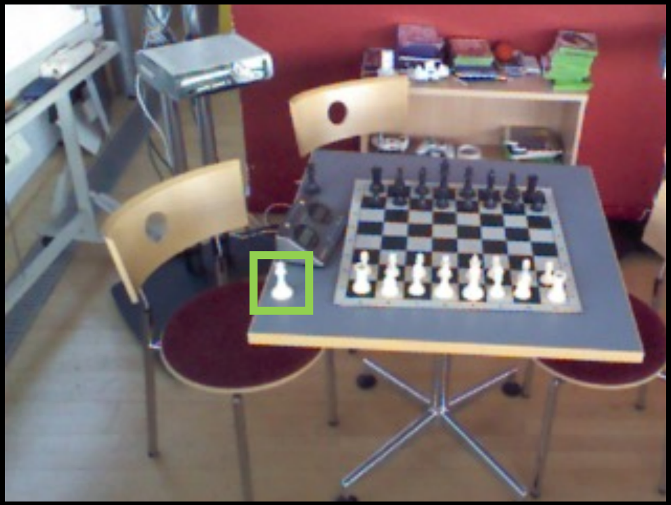
LENS: Localization enhanced by NeRF synthesis, Moreau et al., CoRL 2021

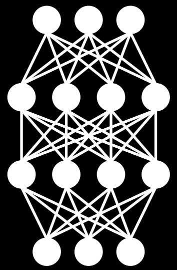
DFNet: Enhance Absolute Pose Regression with Direct Feature Matching, Chen et al., ECCV 2022

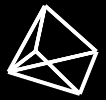
data augmentation via view synthesis


Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
PoseNet15
(45cm)
Spatial LSTM
(31cm)
DenseVLAD (26cm)
PoseNet17
(23cm)
MapNet
(21cm)
Atloc
(20cm)
MS-Transformer
(18cm)
LENS
(8cm)
DFNet
(8cm)
ActiveSearch (4cm)
Visual Camera Re-Localization From RGB and RGB-D Images Using DSAC
Brachmann et al., TPAMI 2021
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

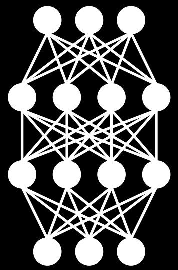
Training Image



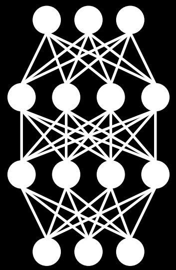


$\hat{h}$ 

Test Images



? ? ?



? ? ?

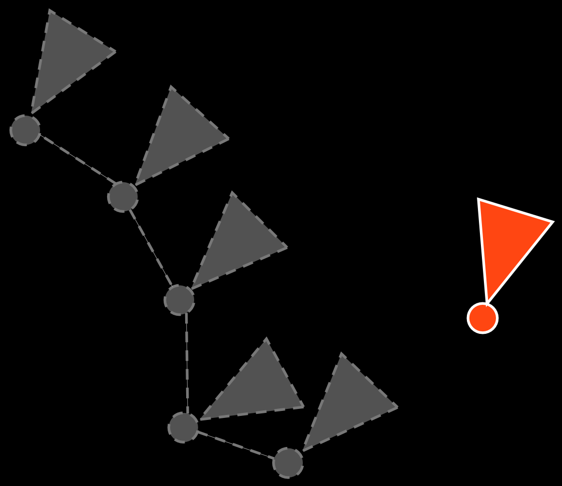
Global methods (APR) need to **extrapolate**.

Local methods (AS, HLoc, DSAC*, NeRF) need to be **invariant**.

Recommended Reading: "Understanding the Limitations of CNN-based Absolute Camera Pose Regression", Sattler et al., CVPR'19

Absolute Pose Regression

- Can be fast at query time
- Slow at mapping time
- Moderate accuracy (but there is progress)



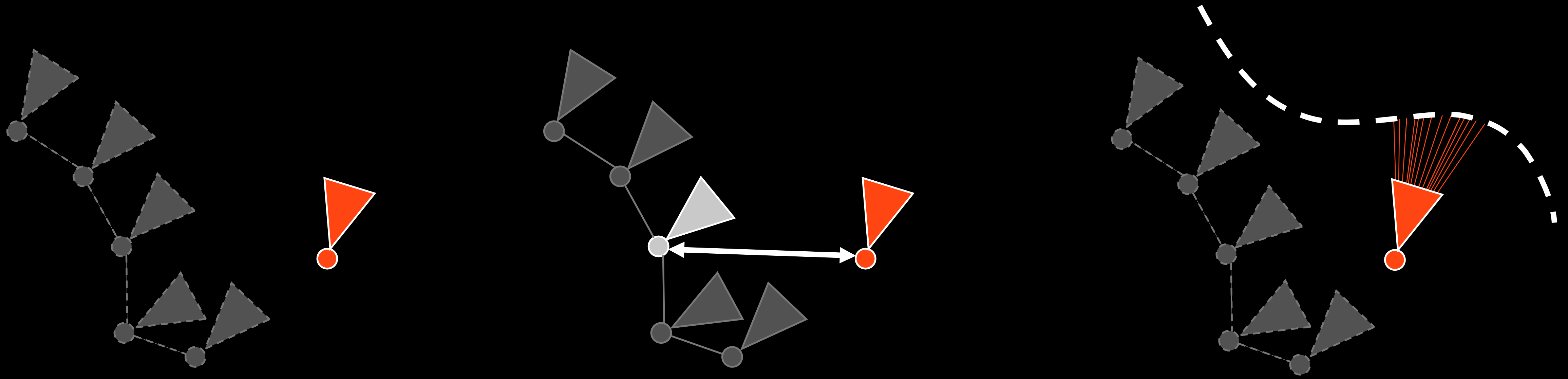
Regression

Pose Regression

Absolute Pose Regression

Relative Pose Regression

Correspondence Regression
(aka Scene Coordinate Regression)



Regression

Pose Regression

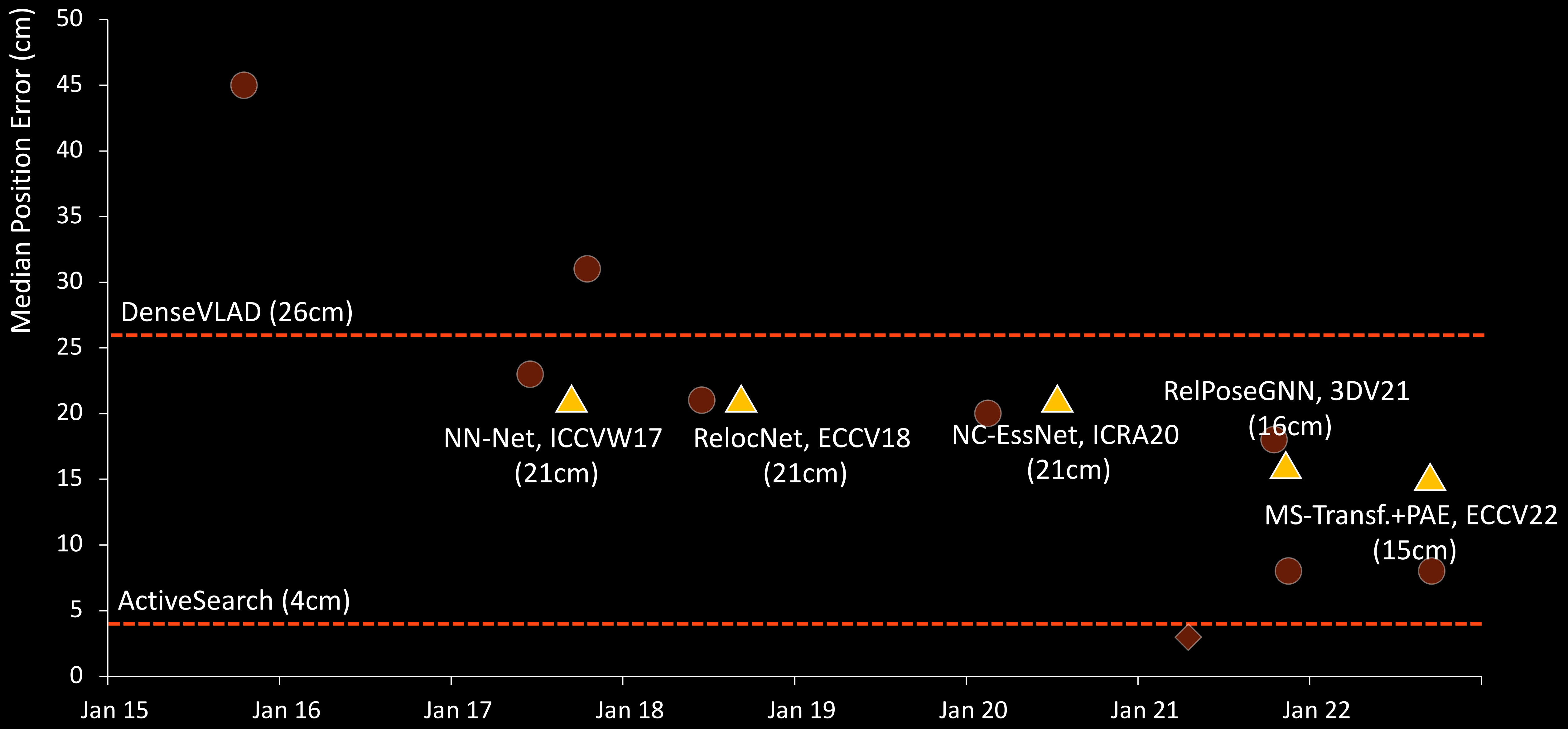
Absolute Pose Regression

Relative Pose Regression

Correspondence Regression
(aka Scene Coordinate Regression)

Scene-specific

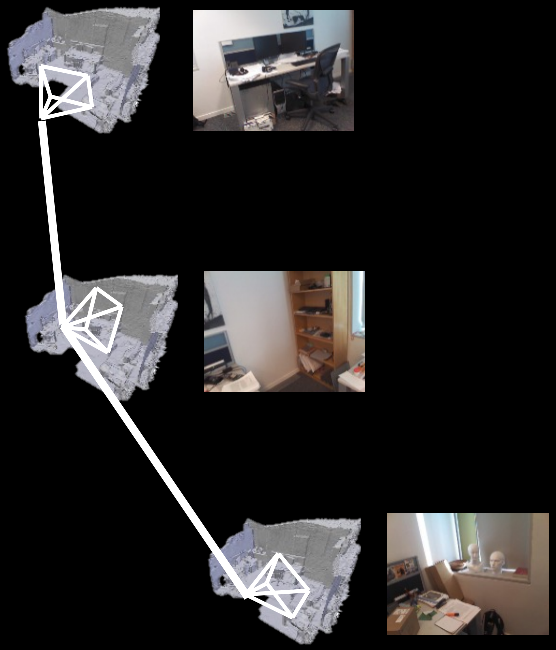
Scene-agnostic


Median Position Error (cm)
50
45
40
35
30
25
20
15
10
5
0
DenseVLAD (26cm)
NN-Net, ICCVW17
(21cm)
RelocNet, ECCV18
(21cm)
NC-EssNet, ICRA20
(21cm)
RelPoseGNN, 3DV21
(16cm)
MS-Transf.+PAE, ECCV22
(15cm)
ActiveSearch (4cm)
Jan 15
Jan 16
Jan 17
Jan 18
Jan 19
Jan 20
Jan 21
Jan 22

Preparation	Mapping	Re-Localisation
Scene-Agnostic Training	Scene-Specific Training	Evaluation

Pre-Train Image Retrieval

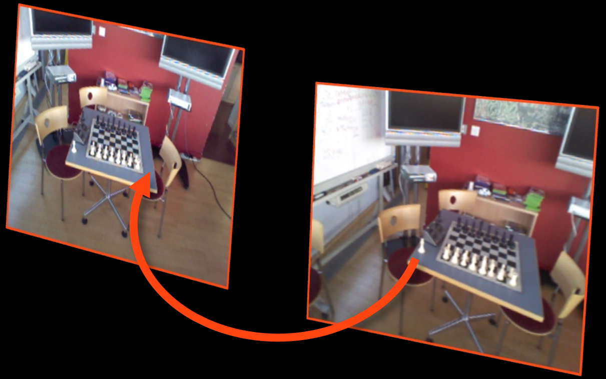
Obtain Posed Images
Build Retrieval Index



Retrieve NN
Refine pose

RelocNet

Pre-Train Relative Pose Regressor



Mapping Time: Very Low

„RelocNet: Continuous Metric Learning Relocalisation using Neural Nets”, Balntas et al., ECCV’18

Preparation	**Mapping**	**Re-Localisation**
Scene-Agnostic Training	**Scene-Specific Training**	**Evaluation**

Pre-Train Image Retrieval

RelocNet

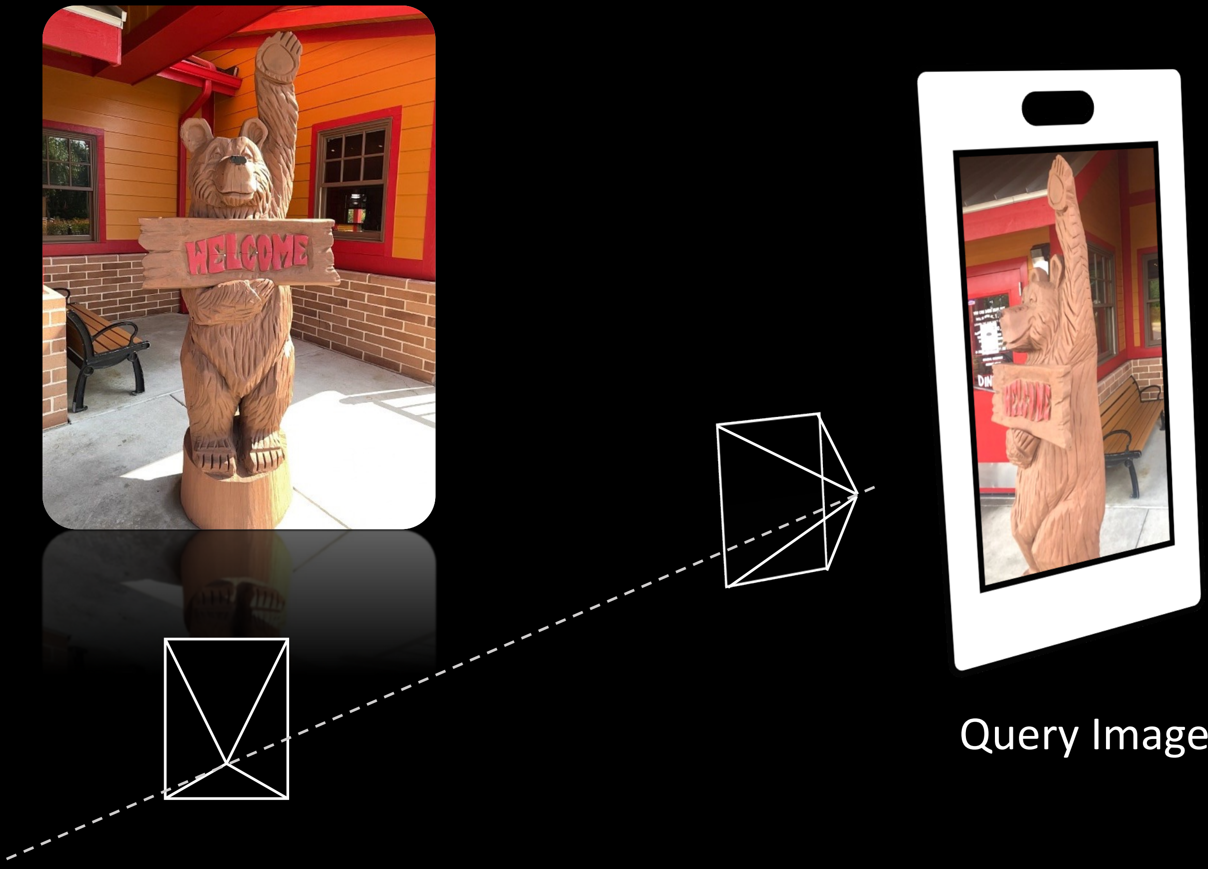
Pre-Train Relative Pose Regressor

Obtain Posed Images
Build Retrieval Index

Retrieve NN
Refine pose

„RelocNet: Continuous Metric Learning Relocalisation using Neural Nets", Balntas et al., ECCV'18

Reference Image



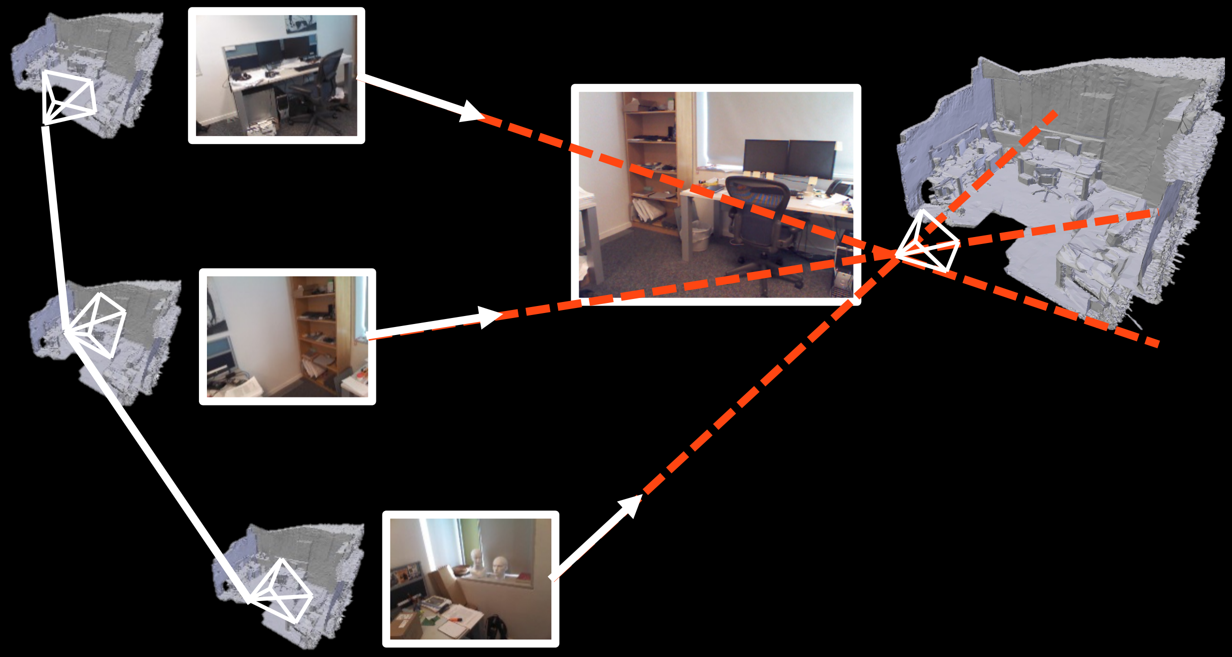
Query Image

RelocNet
Median error on 7Scenes

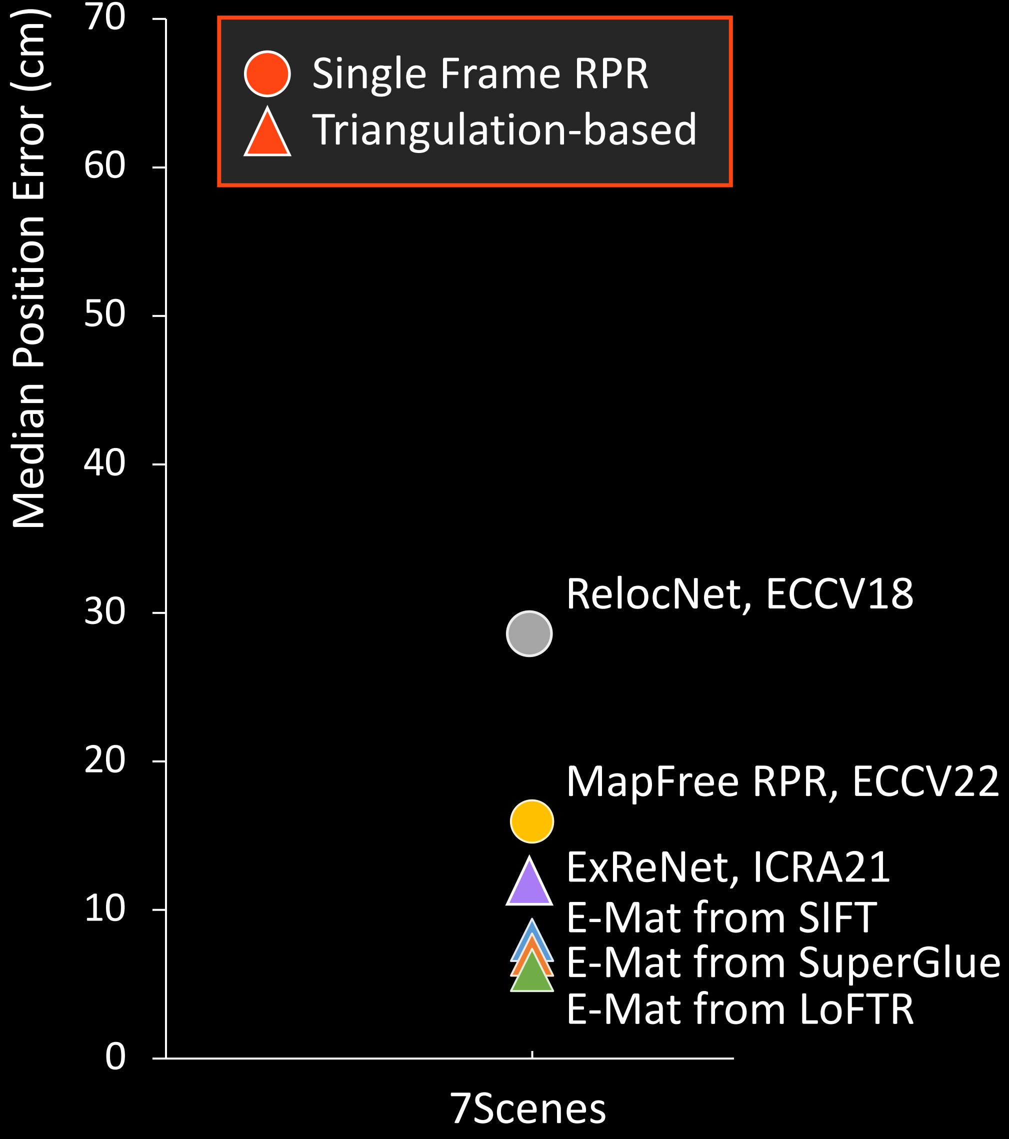
Train on 7Scenes: 21cm
Train on ScanNet: 29cm

RelocNet: Continuous Metric Learning Relocalisation using Neural Nets
Balntas et al., ECCV 2018

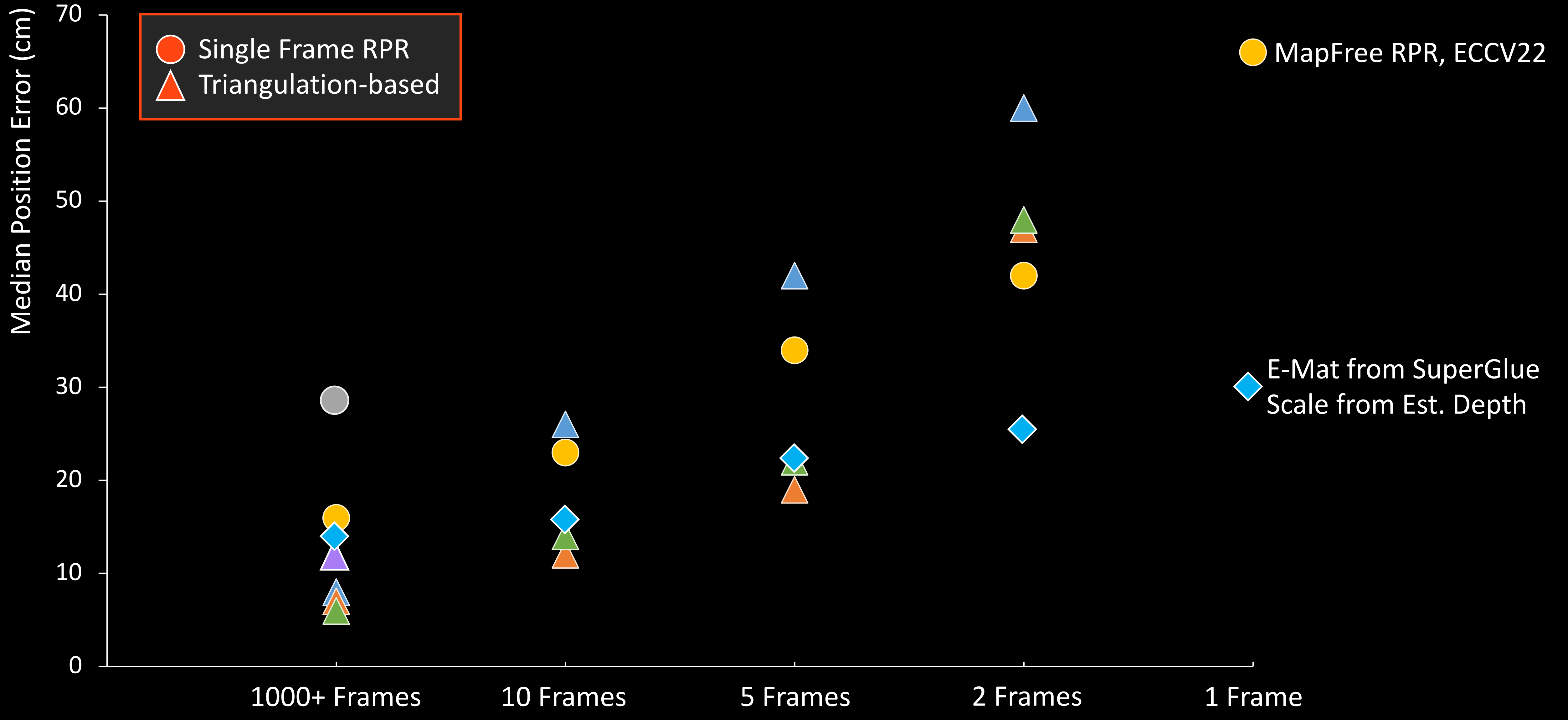
Preparation **Scene-Agnostic Training**	**Mapping** **Scene-Specific Training**	**Re-Localisation** **Evaluation**
Pre-Train Image Retrieval Pre-Train Relative Pose Regressor	Obtain Posed Images Build Retrieval Index	Retrieve NN Refine pose



"To Learn or Not to Learn: Visual Localization from Essential Matrices", Zhou at al., ICRA'20



Data from Arnold et al., "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22



Data from Arnold et al., "Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22

Preparation	Mapping	Re-Localisation
Scene-Agnostic Training	Scene-Specific Training	Evaluation

Pre-Train Image Retrieval

Shoot Reference Image

Mapping Time: Zero

Estimate Relative Pose

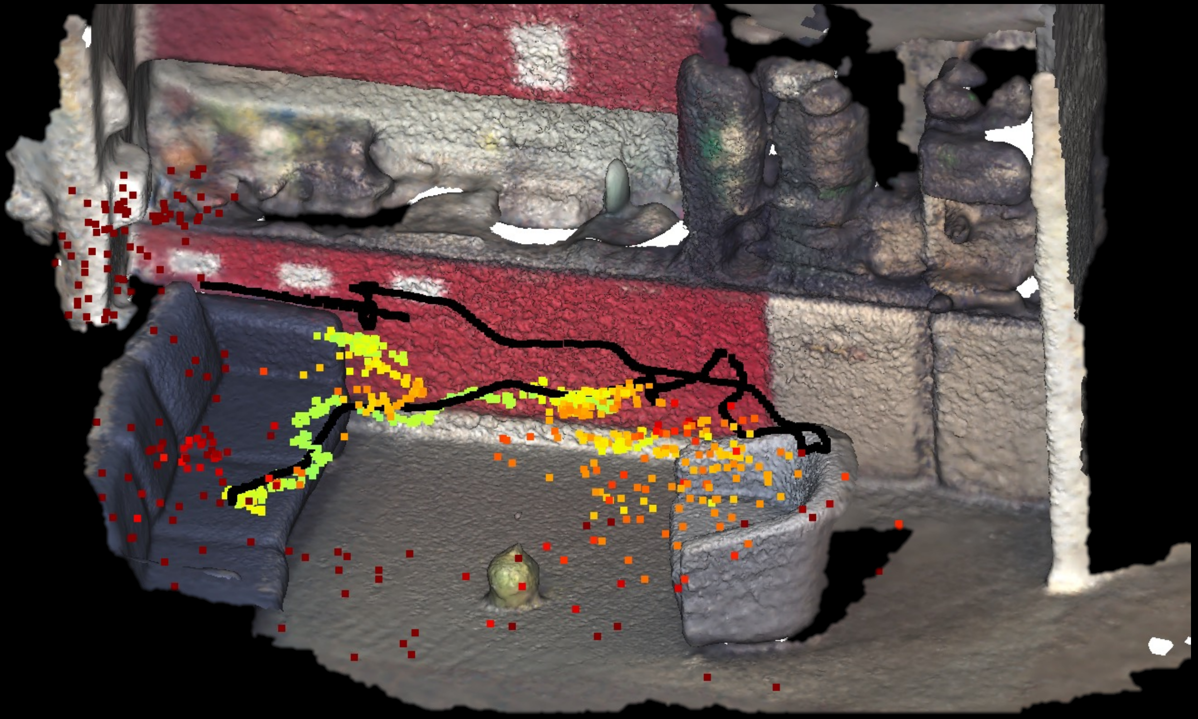
Pre-Train Relative Pose Regressor

"Map-free Visual Relocalization: Metric Pose Relative to a Single Image", ECCV22

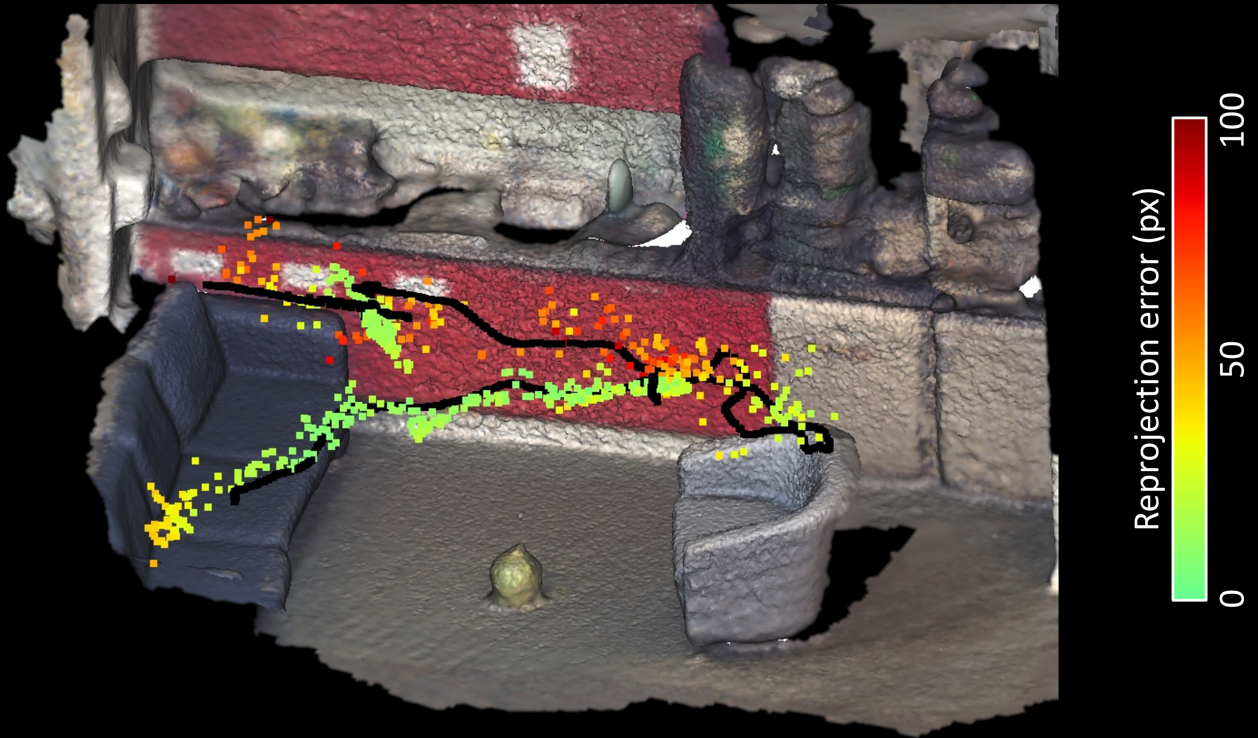
Reference Image

Query Images

MapFree RPR, ECCV22



E-Mat from SuperGlue, Scale from Est. Depth



Map-Free Relocaliasation Dataset

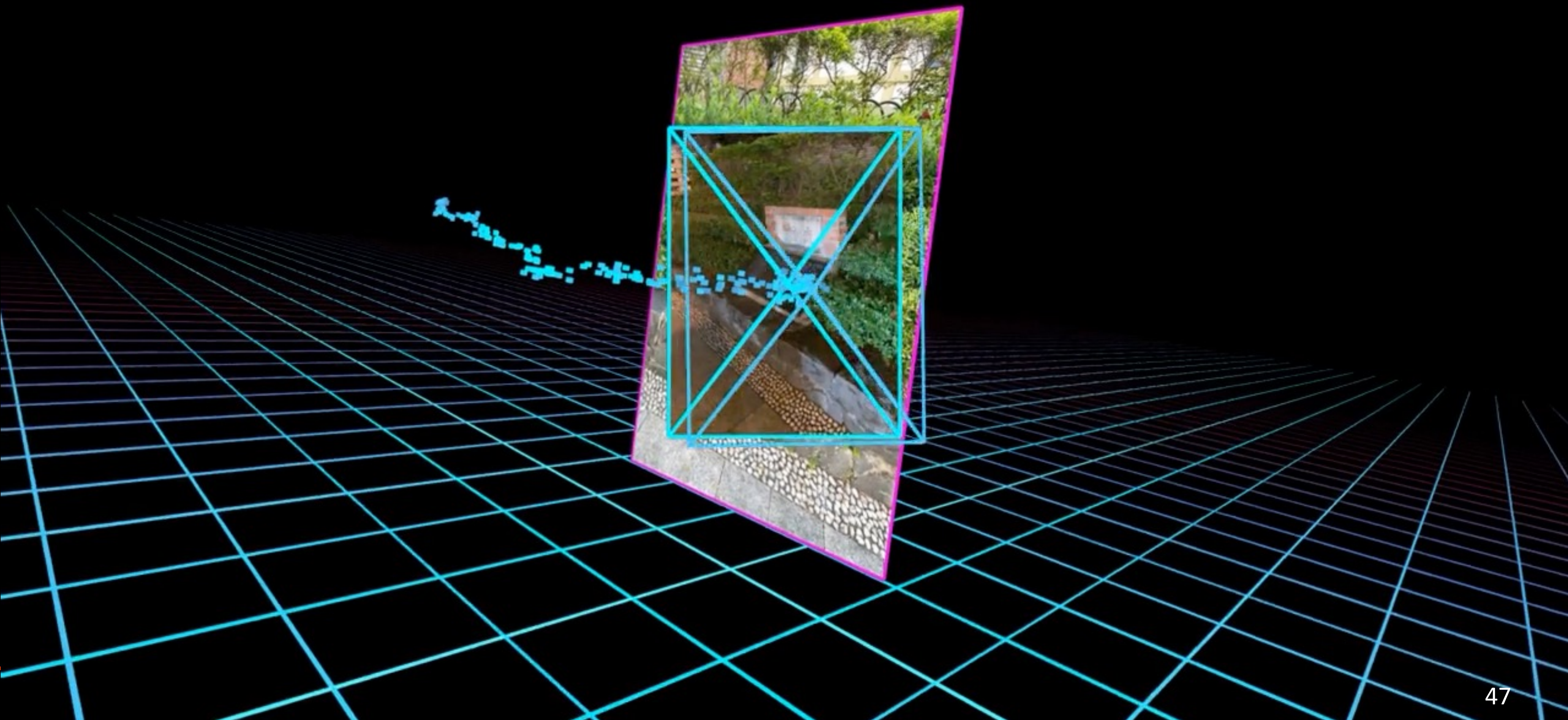
Validation
65 Scenes

Test
130 Scenes

Training
460 Scenes

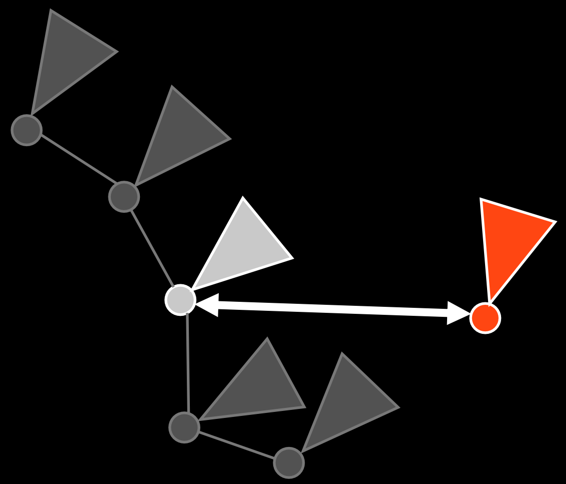
Dataset, Baselines, Evaluation Code, Leaderboard

https://research.nianticlabs.com/mapfree-reloc-benchmark



Reference Frame
Query Ground Truth
SuperGlue + Est.Depth
Relative Pose Regression

Relative Pose Regression



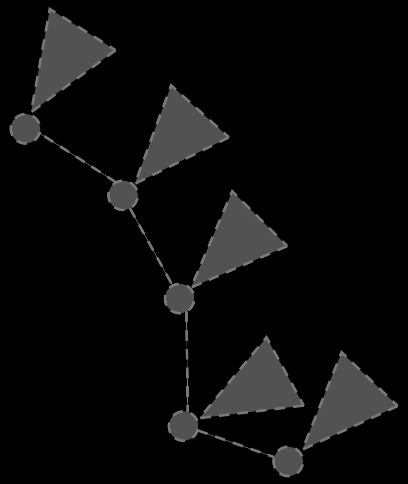
- Low or no mapping costs
- Scale ambiguity is key challenge
- New dataset and benchmark to measure and drive progress:

  Map-Free Relocalisation

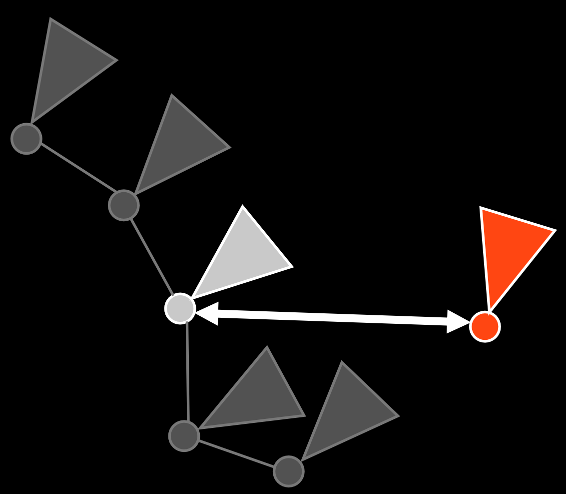
Regression

Pose Regression

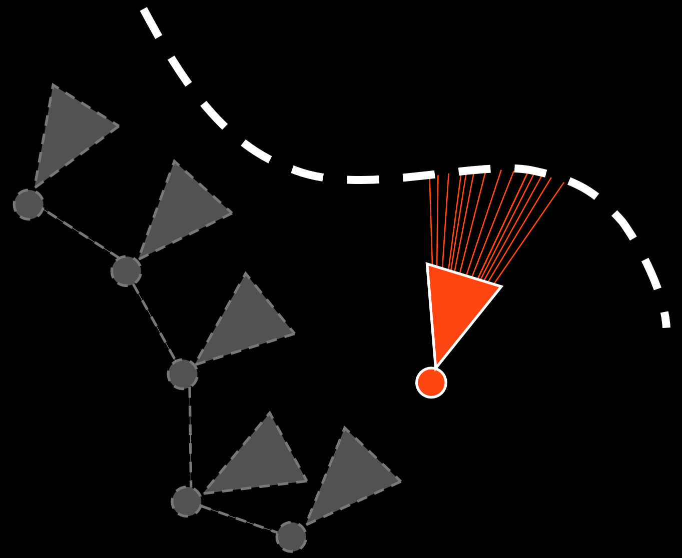
Absolute Pose Regression



Relative Pose Regression



Correspondence Regression (aka Scene Coordinate Regression)



Query Image I	Image Retrieval	Feature Extraction	Feature Matching	RANSAC: Estimate Poses (PnP), Score Poses (Inliers)	6D Pose $\mathbf{h}$

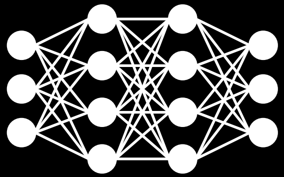
Query Image I	Image Retrieval	Correspondence Regression	RANSAC: Estimate Poses (PnP), Score Poses (Inliers)	6D Pose $\mathbf{h}$

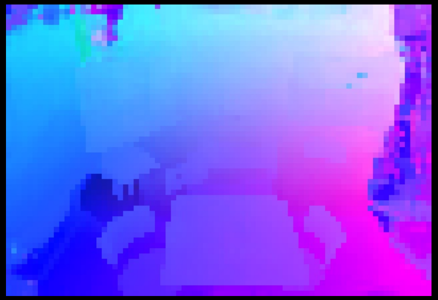
Query Image

Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR'13

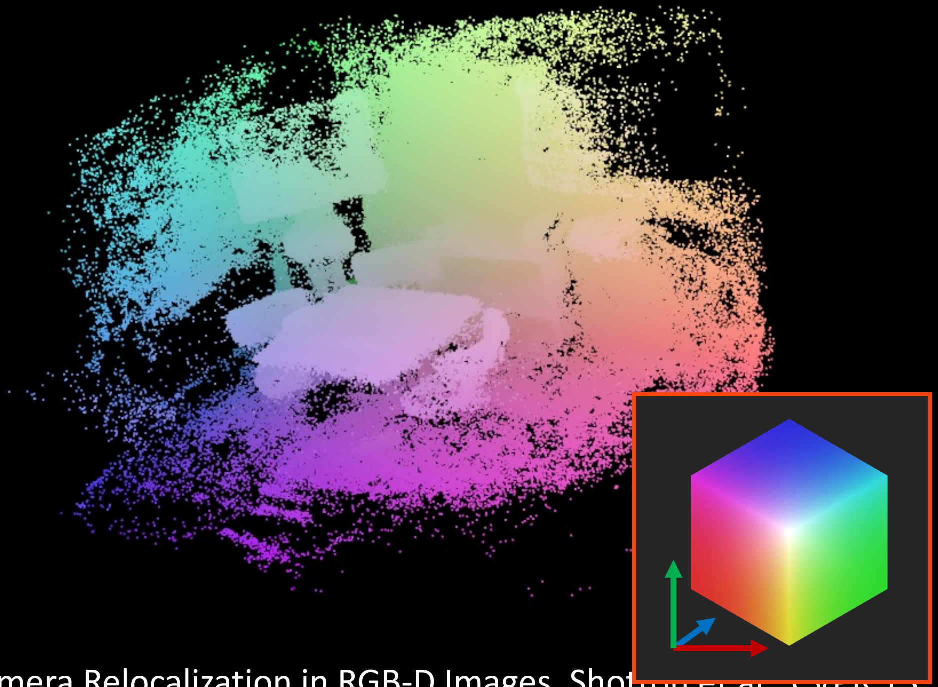
Query Image I	Image Retrieval	Feature Extraction	Feature Matching	RANSAC	Estimate Poses (PnP) Score Poses (Inliers)	6D Pose $\mathbf{h}$
Query Image I	Image Retrieval	Correspondence Regression		RANSAC	Estimate Poses (PnP) Score Poses (Inliers)	6D Pose $\mathbf{h}$

Query Image





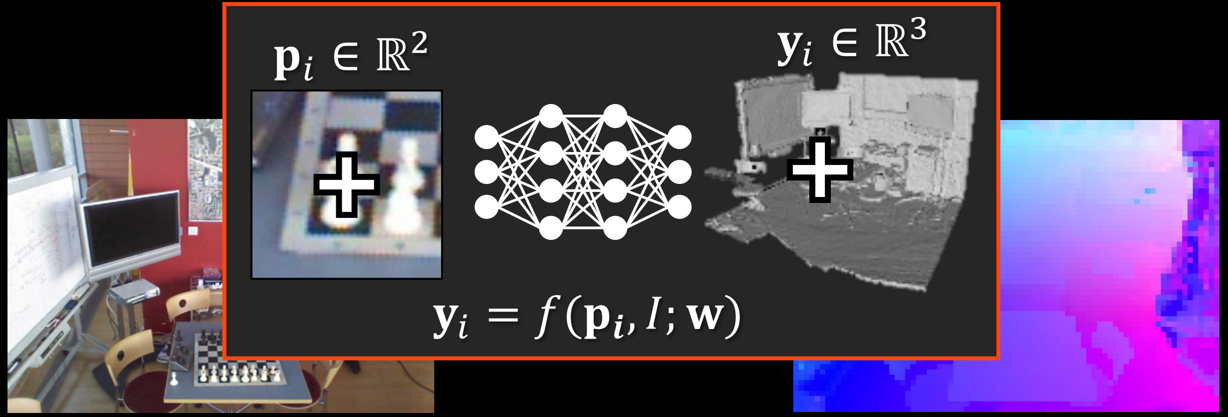
Scene Coordinate Prediction



Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR'13

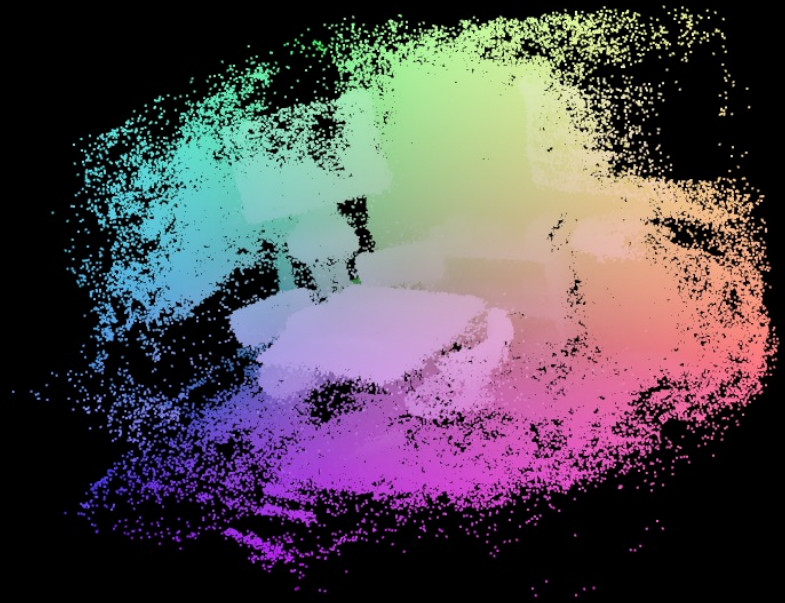
Query Image I | Image Retrieval | Feature Extraction | Feature Matching | RANSAC (Estimate Poses (PnP), Score Poses (Inliers)) | 6D Pose $\mathbf{h}$

Query Image I | Image Retrieval | Correspondence Regression | RANSAC (Estimate Poses (PnP), Score Poses (Inliers)) | 6D Pose $\mathbf{h}$



Query Image

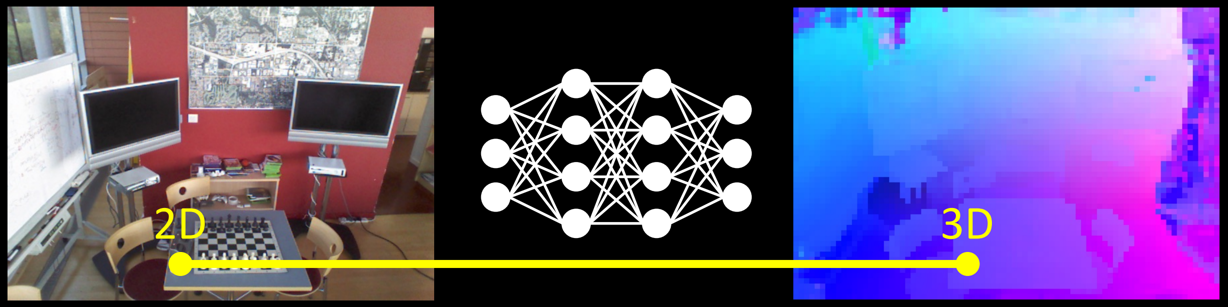
Scene Coordinate Prediction



Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR'13

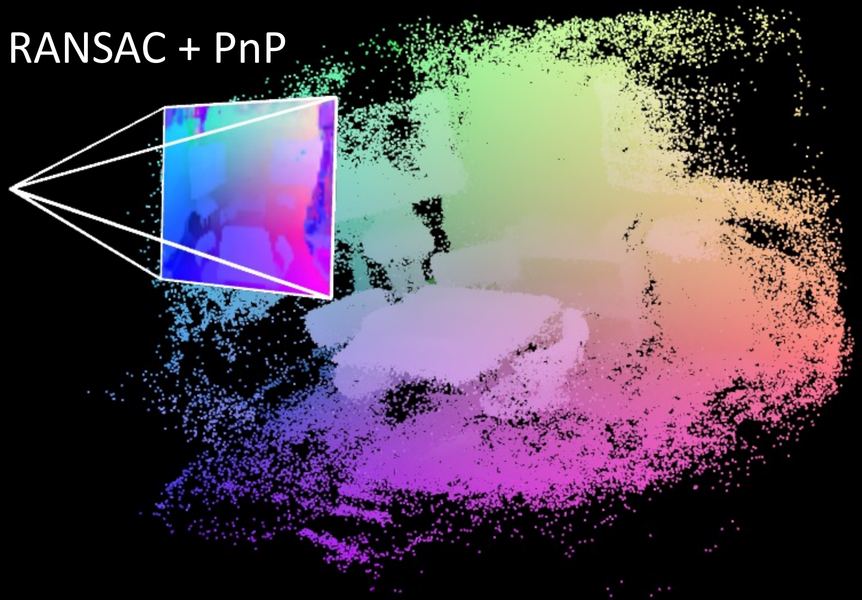
Query Image I → Image Retrieval → Feature Extraction → Feature Matching → RANSAC (Estimate Poses (PnP), Score Poses (Inliers)) → 6D Pose $\mathbf{h}$

Query Image I → Image Retrieval → Correspondence Regression → RANSAC (Estimate Poses (PnP), Score Poses (Inliers)) → 6D Pose $\mathbf{h}$



Query Image

Scene Coordinate Prediction

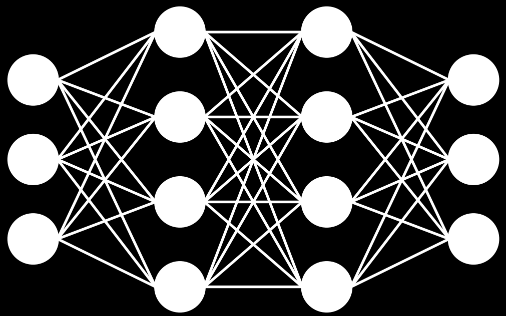


Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR’13

Case I: Mapping images are RGB-D or 3D model is available

$$\ell_{L_1}(\mathbf{y}, \mathbf{y}^*) = \|\mathbf{y} - \mathbf{y}^*\|$$

Mapping Image



Scene Coordinate Prediction

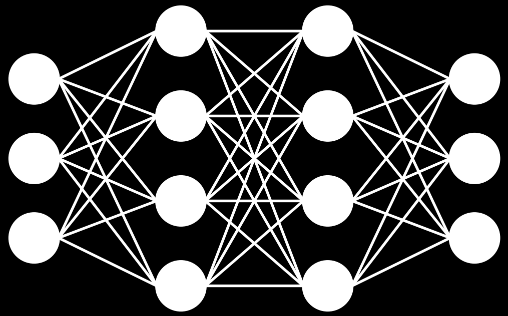
Scene Coordinate Ground Truth

Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC, Brachmann and Rother, TPAMI'21

Case I: Mapping images are RGB-D or 3D model is available

$$\ell_{L_1+\pi}(\mathbf{y}, \mathbf{y}^*, \mathbf{h}^*) = \begin{cases} \|\mathbf{y} - \mathbf{y}^*\|, & \|\mathbf{y} - \mathbf{y}^*\| > 0.1\text{m} \\ \|\mathbf{p} - \pi(\mathbf{y}, \mathbf{h}^*)\|, & \|\mathbf{y} - \mathbf{y}^*\| < 0.1\text{m} \end{cases}$$

Mapping Image



Scene Coordinate Prediction

Scene Coordinate Ground Truth

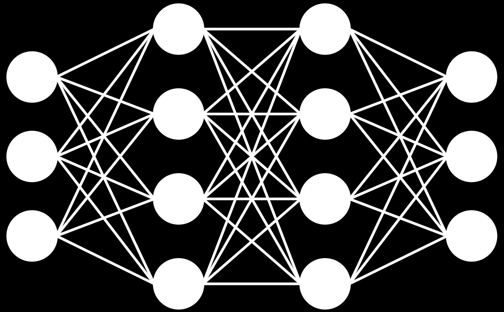
Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC, Brachmann and Rother, TPAMI'21

Case II: Mapping images are RGB

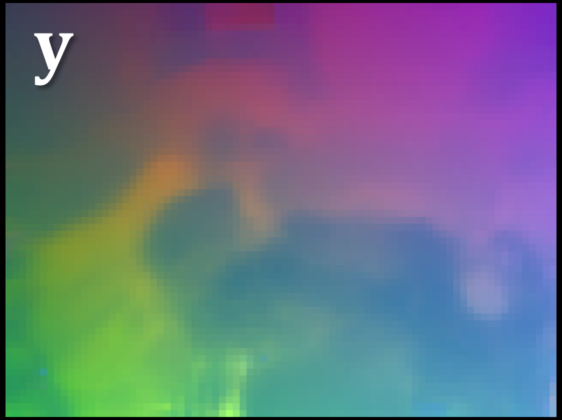
$$\mathcal{V} = \left\{ \mathbf{y} \;\middle|\; \begin{array}{c} \text{in front of camera} \\ \text{below maximum distance} \\ \text{below maximum reprojection error} \end{array} \right\}$$

$$\ell_{L_1+\pi}(\mathbf{y}, \mathbf{h}^*) = \begin{cases} \|\mathbf{p} - \pi(\mathbf{y}, \mathbf{h}^*)\|, & \mathbf{y} \in \mathcal{V} \\ \|\mathbf{y} - \bar{\mathbf{y}}\|, & \text{otherwise} \end{cases}$$

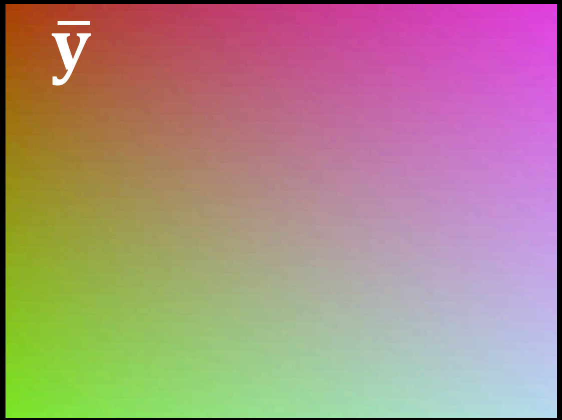
Mapping Image



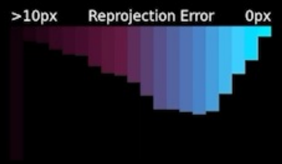
Scene Coordinate Prediction

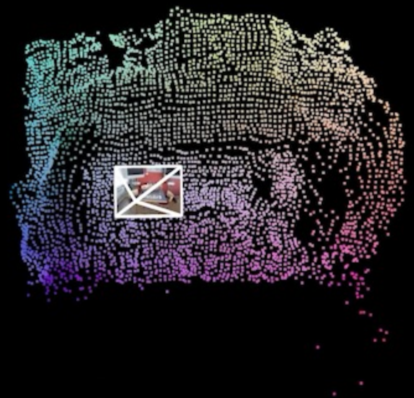


Scene Coordinate Heuristic

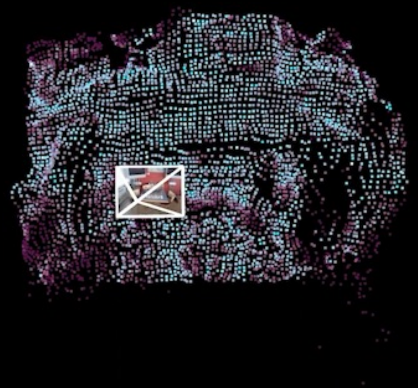


Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC, Brachmann and Rother, TPAMI'21



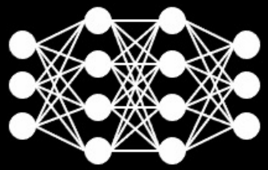


Scene Coordinates



Reprojection Loss

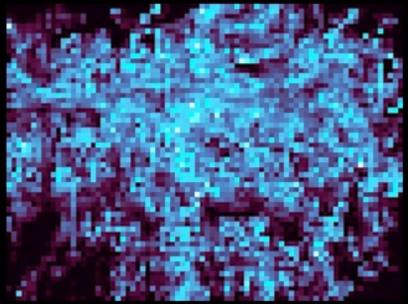
Mapping Image



Network

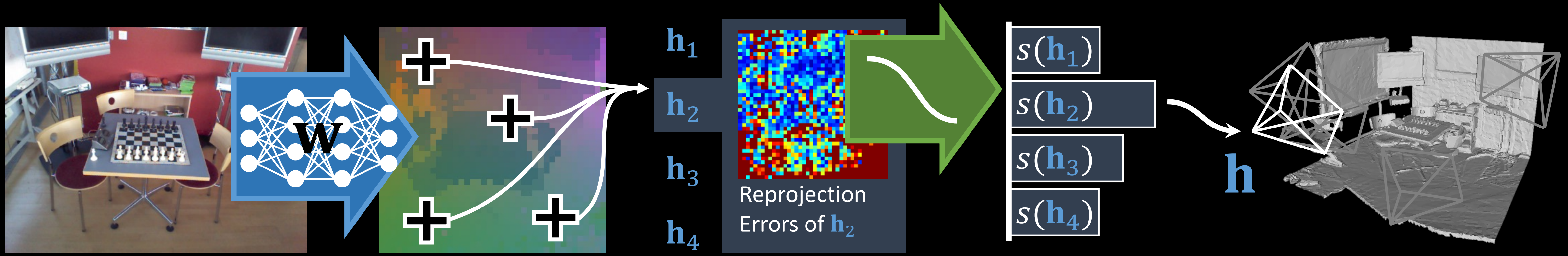
Scene Coordinates y

π



Reprojection Loss ℓ

Image I | Scene Coordinate Regression (CNN) | RANSAC: Estimate Poses (PnP), Score Poses (Inliers) | 6D Pose $\mathbf{h}$



Input RGB | Scene Coordinate Regression | Hypothesis Sampling | Scoring | Hypothesis Selection | Result

Differentiating PnP [DSAC++]:

$$\mathbf{h} = \operatorname*{argmin}_{\mathbf{h}'} \sum_{i=0}^{4} \|\mathbf{p}_i - \pi(\mathbf{y}_i, \mathbf{h}')\|^2$$

Soft Inlier Counting [DSAC++] :

$$s(\mathbf{h}) = \sum_i \operatorname{sig}(\tau - \beta \|\mathbf{p}_i - \pi(\mathbf{y}_i, \mathbf{h})\|)$$

argmax Selection

$$\mathbf{h} = \operatorname*{argmax}_{\mathbf{h}_j} s(\mathbf{h}_j)$$

[DSAC] "DSAC - Differentiable RANSAC for camera localization", Brachmann et al., CVPR'17
[DSAC++] "Learning less is more - 6D camera localization via 3D surface regression", Brachmann and Rother, CVPR'18

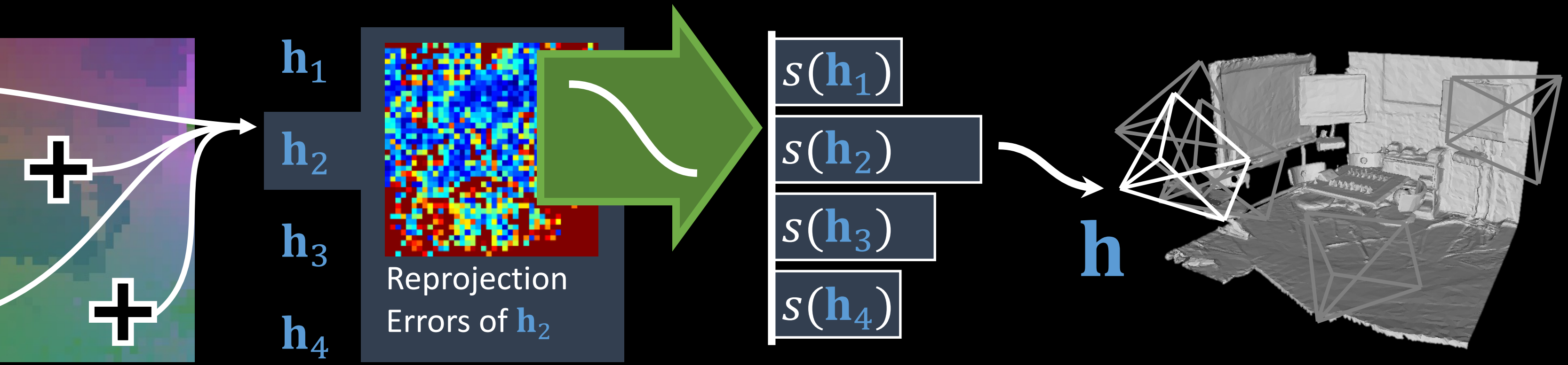
inate Regression
CNN)

RANSAC
- Estimate Poses (PnP)
- Score Poses (Inliers)

6D Pose **h**

DSAC Learning objective:

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{j\sim P(j;\mathbf{w})}\left[\ell(\mathbf{h}_j, \mathbf{h}^*)\right]$$



ession Hypothesis Sampling Scoring Hypothesis Selection Result

$|^2$

Soft Inlier Counting [DSAC++] :

$$s(\mathbf{h}) = \sum_i \mathrm{sig}(\tau - \beta\|\mathbf{p}_i - \pi(\mathbf{y}_i, \mathbf{h})\|)$$

argmax Selection

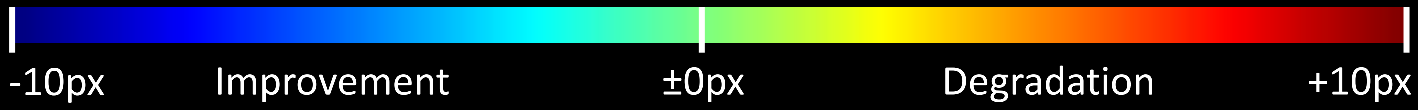
$$\mathbf{h} = \underset{\mathbf{h}_j}{\operatorname{argmax}}\, s(\mathbf{h}_j)$$

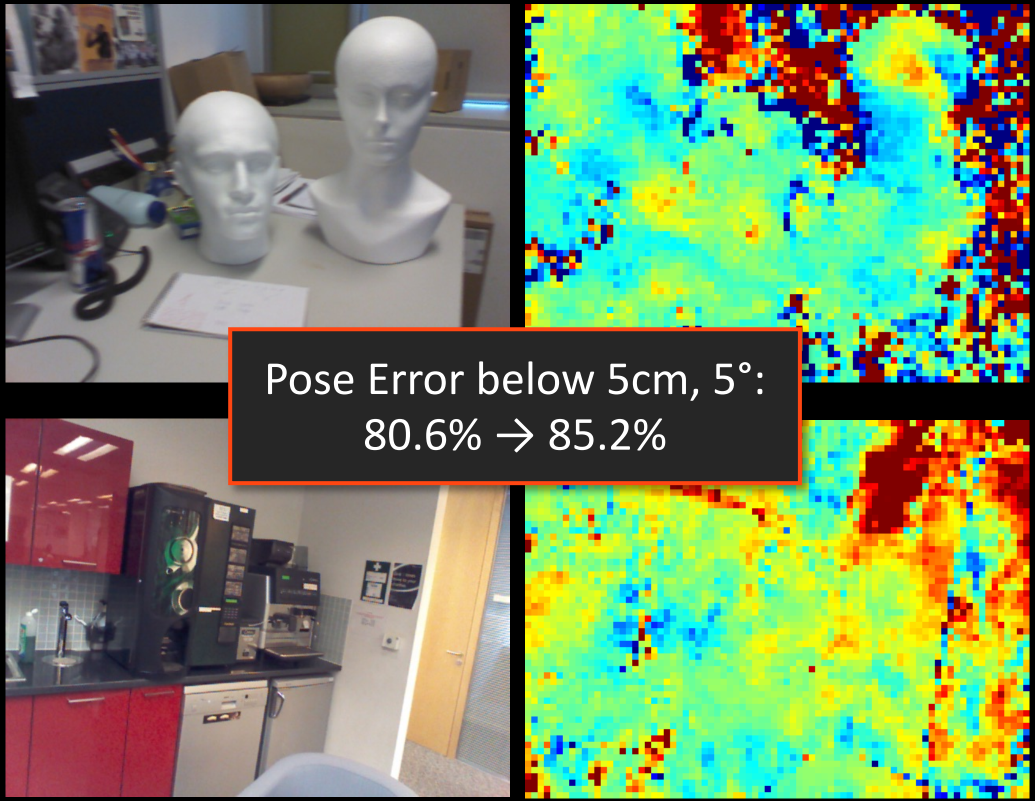
Probabilistic Selection [DSAC]

$$\hat{\mathbf{h}} = \mathbf{h}_j, \text{where } j \sim \frac{\exp(s(\mathbf{h}_j))}{\sum_k \exp(s(\mathbf{h}_k))}$$

[DSAC] "DSAC - Differentiable RANSAC for camera localization", Brachmann et al., CVPR'17
[DSAC++] "Learning less is more - 6D camera localization via 3D surface regression", Brachmann and Rother, CVPR'18

Comparing reprojection error before and after end-to-end training:



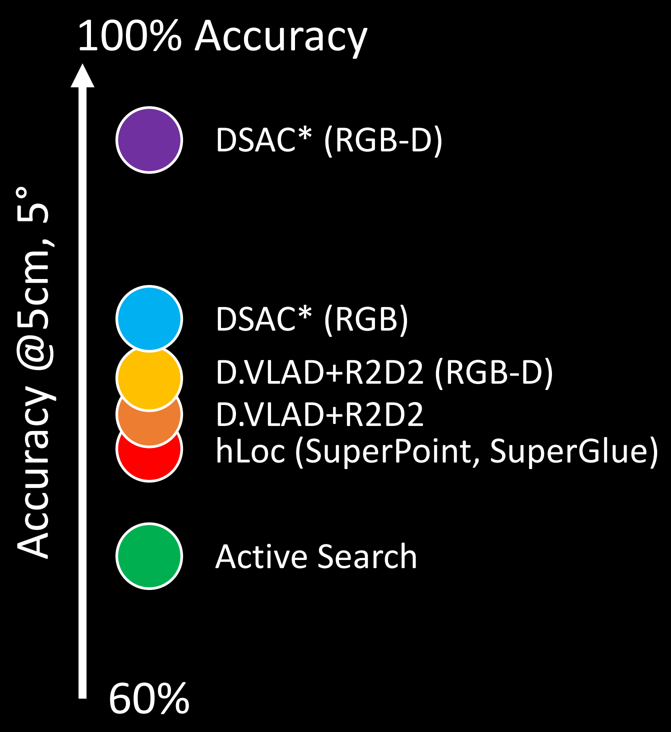


7Scenes Dataset [Sho13]

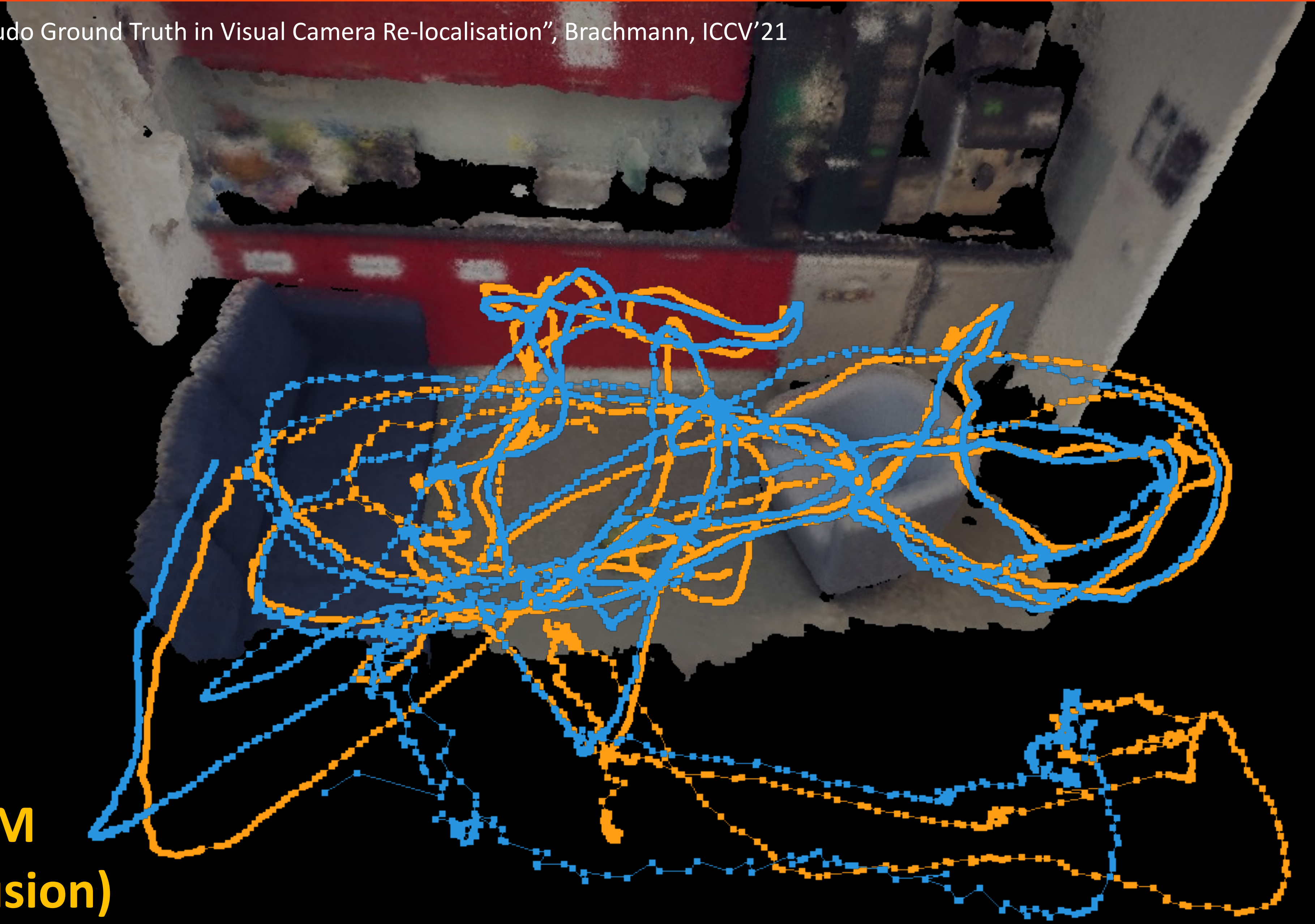
Cambridge Landmarks [Ken15]

[Sho13] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13

[Ken15] “PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Localization”, Kendall et al., ICCV’15


100% Accuracy
Accuracy @5cm, 5°
DSAC* (RGB-D)
DSAC* (RGB)
D.VLAD+R2D2 (RGB-D)
D.VLAD+R2D2
hLoc (SuperPoint, SuperGlue)
Active Search
60%

"On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann, ICCV'21



D-SLAM (Kinect Fusion)

SfM (COLMAP)

“On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation”, Brachmann, ICCV’21

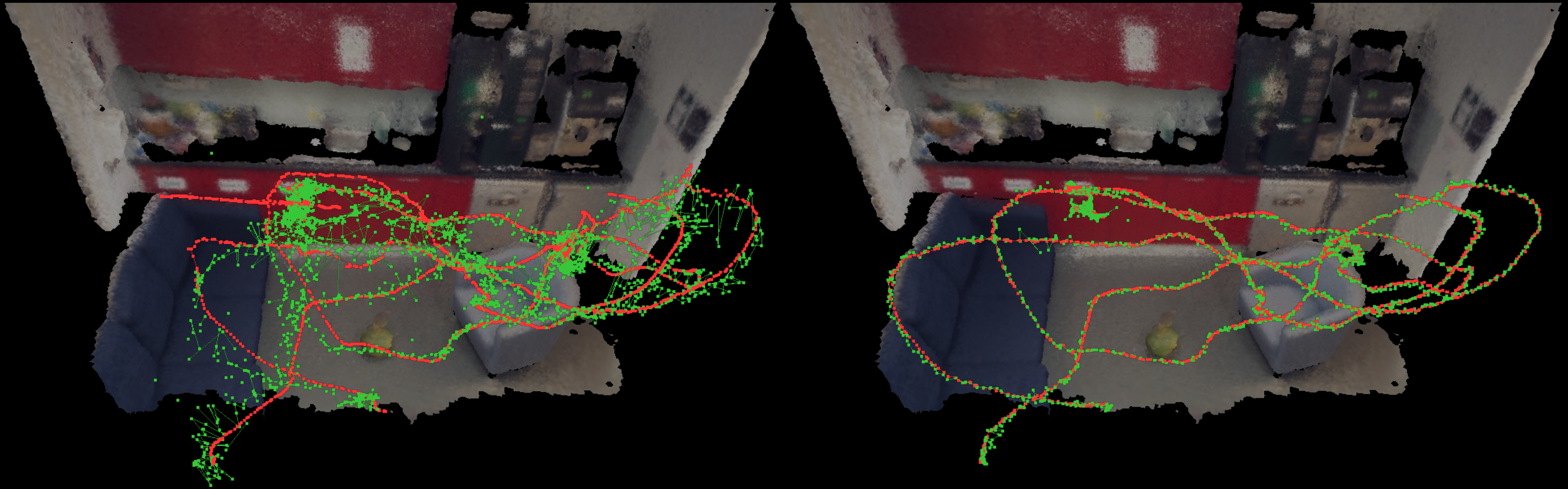
Active Search

D-SLAM (Kinect Fusion) pseudo ground truth

SfM (COLMAP) pseudo ground truth

"On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann, ICCV'21

Active Search

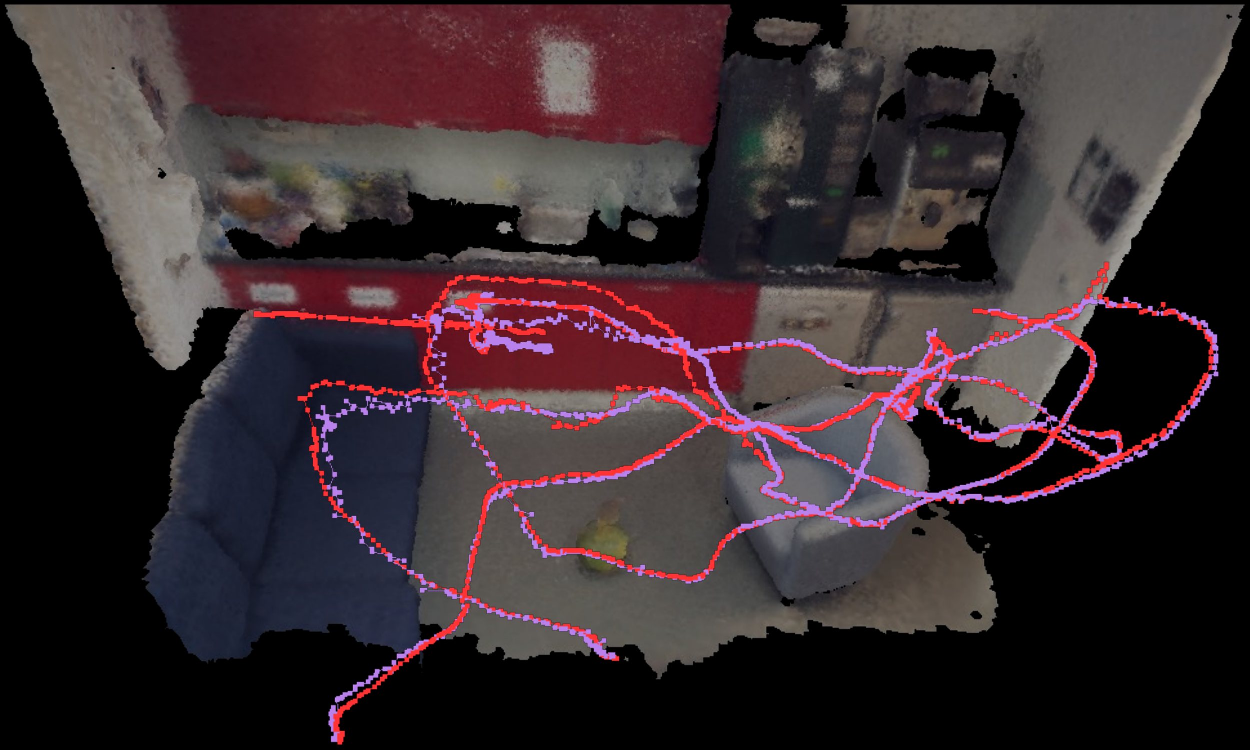


D-SLAM (Kinect Fusion) pseudo ground truth

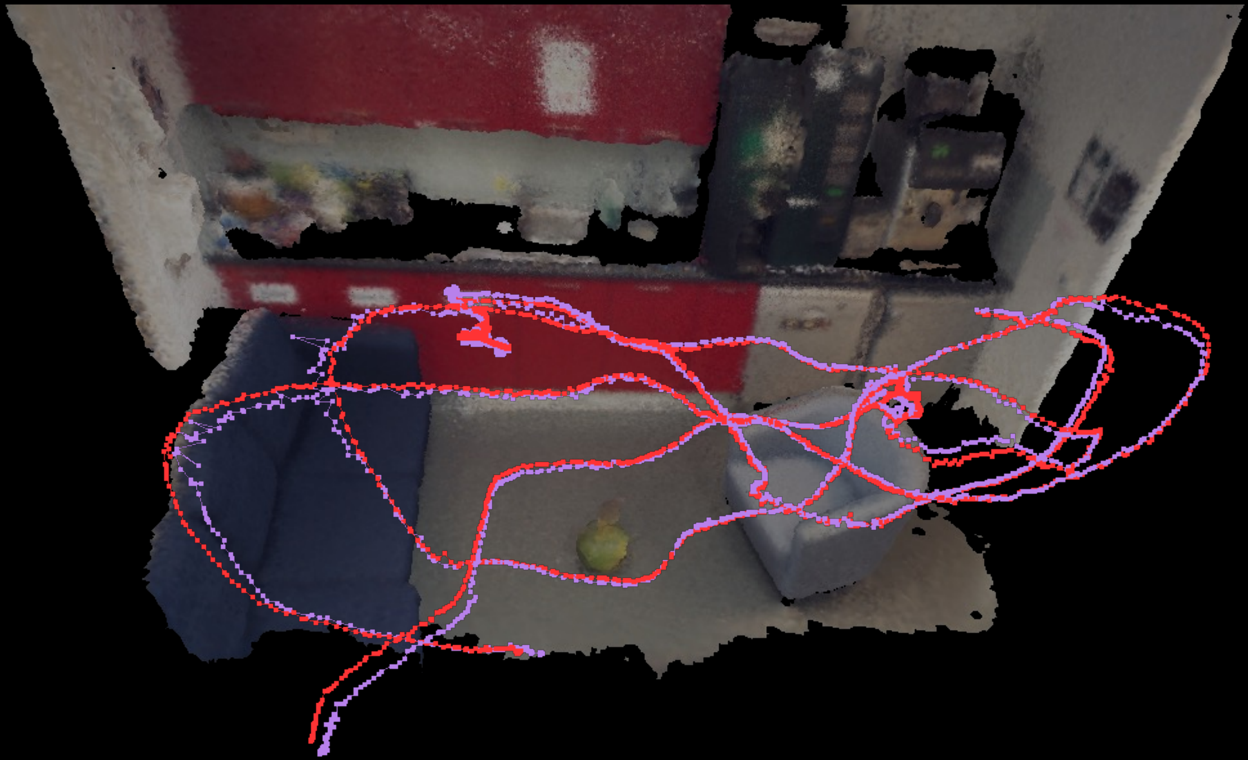
SfM (COLMAP) pseudo ground truth

"On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann, ICCV'21

DSAC* (RGB-D)

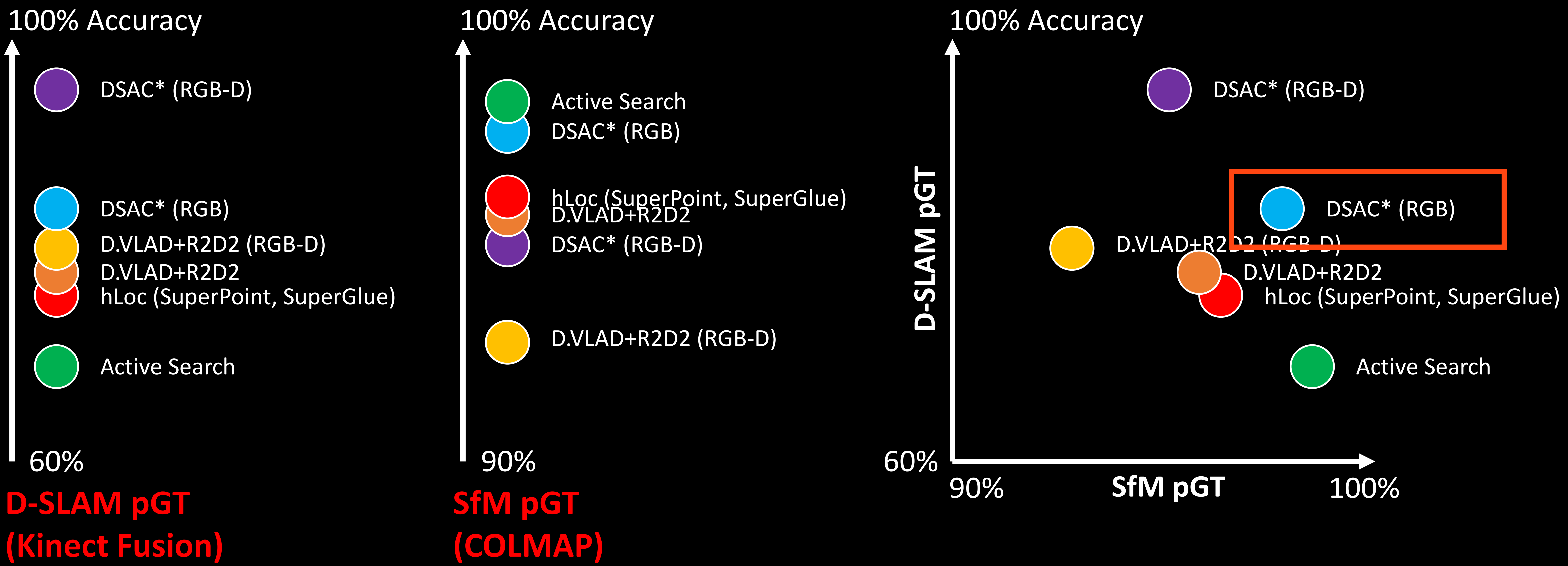


D-SLAM (Kinect Fusion) pseudo ground truth



SfM (COLMAP) pseudo ground truth

E. Brachmann and C. Rother. "Visual camera relocalization from RGB and RGB-D images using DSAC". TPAMI, 2021



"On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann, ICCV'21

Reprojection Error

End-to-end Pose Error

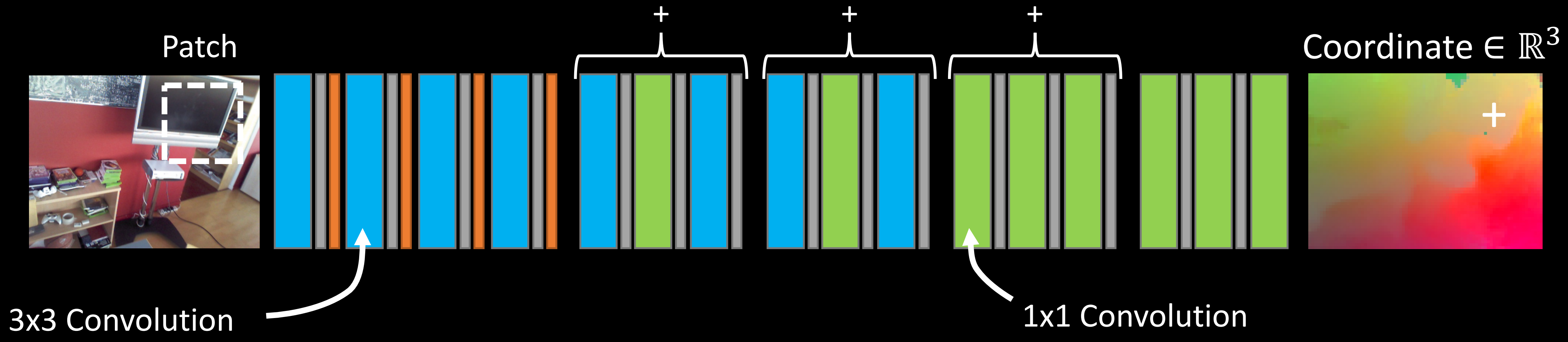
Training Time: 15 hours

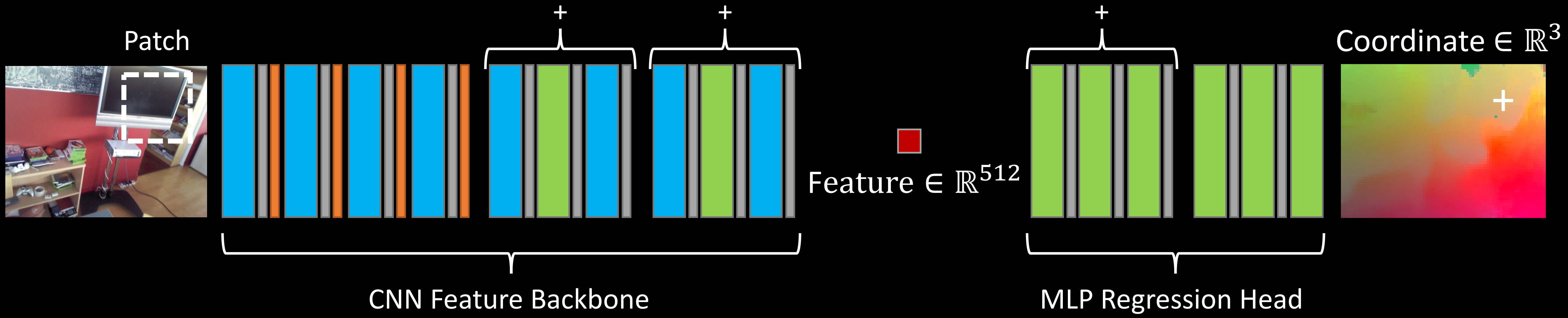
Some Strategies:

- RGB-D mapping, RGB-D queries:
  - Random Forests: <10 minutes
    - Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al., CVPR13
  - On-the-fly Adapatation: realtime
    - On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation, Cavallari et al., CVPR17
    - Let's take this online: Adapting scene coordinate regression network predictions for online RGB-D camera relocalisation, Cavallari et al., 3DV19

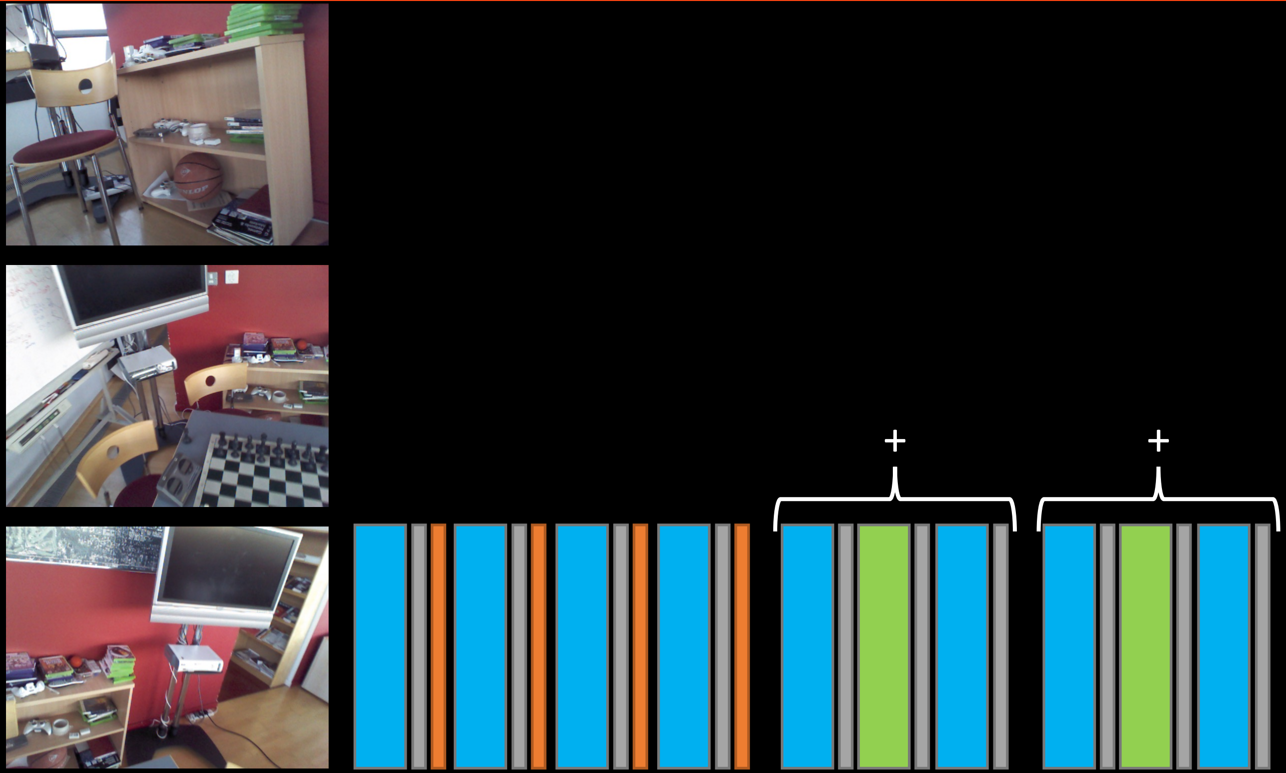
- RGB-D mapping, RGB queries:
  - Random Forests: <10 minutes
    - Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image, Brachmann et al., CVPR16
  - Few Shot Learning, Meta Learning: <5 minutes
    - Visual Localization via Few-Shot Scene Region Classification, Dong et al., 3DV22

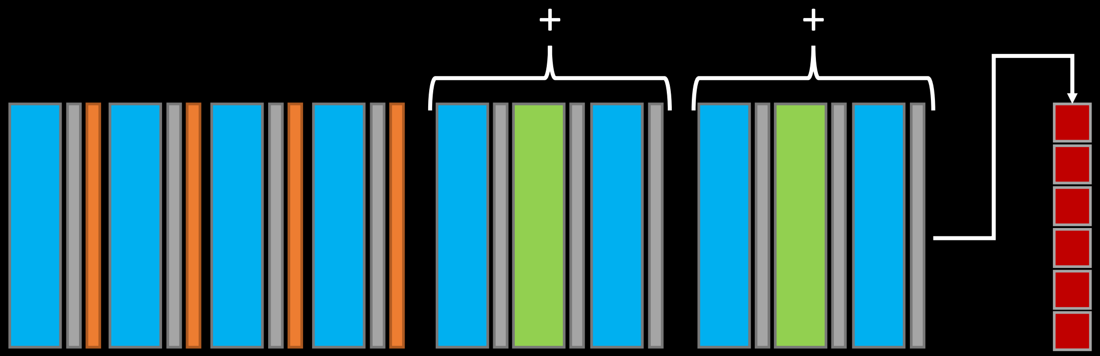
ACE – CVPR23 Highlight – TUE PM 86

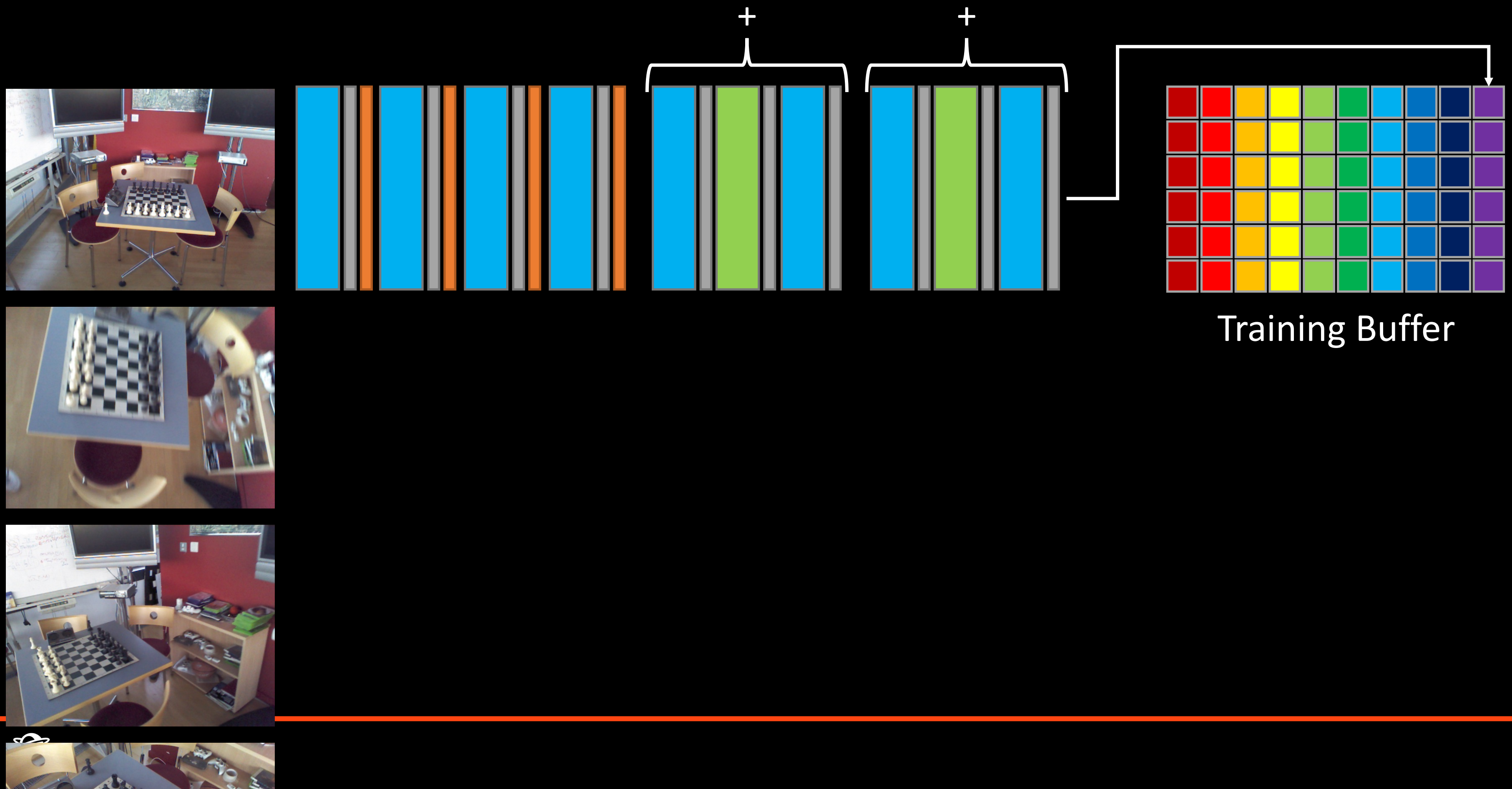


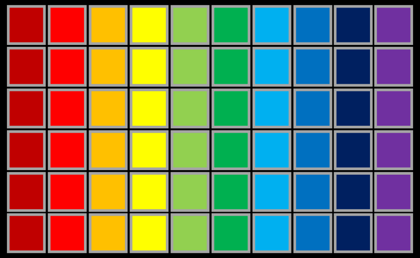


Patch
+
+
+
Coordinate ∈ $\mathbb{R}^3$
Feature ∈ $\mathbb{R}^{512}$
CNN Feature Backbone
MLP Regression Head



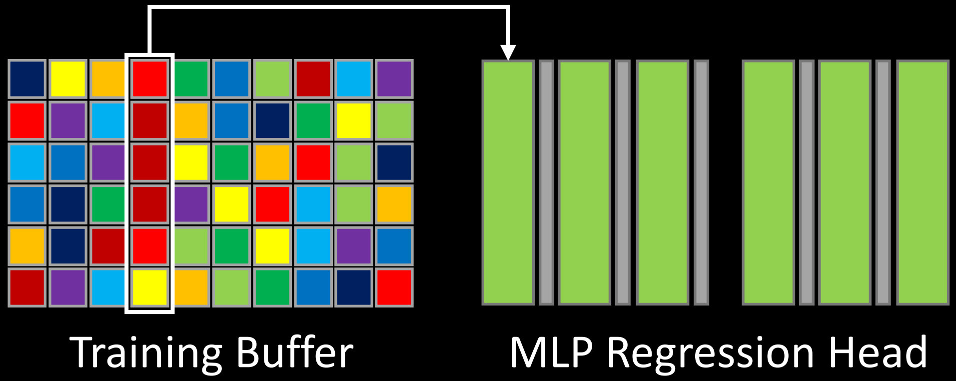


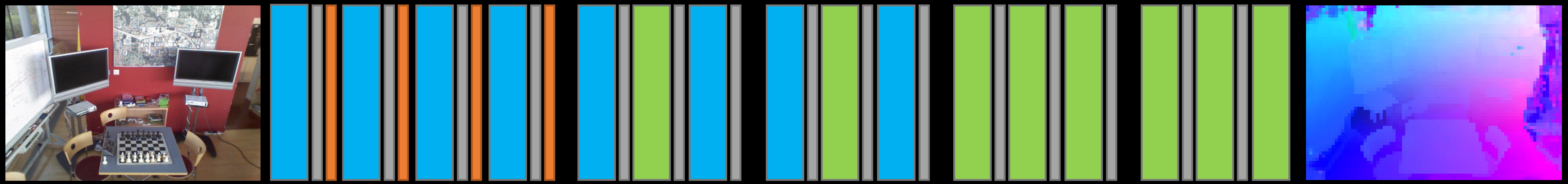

+
+
Training Buffer

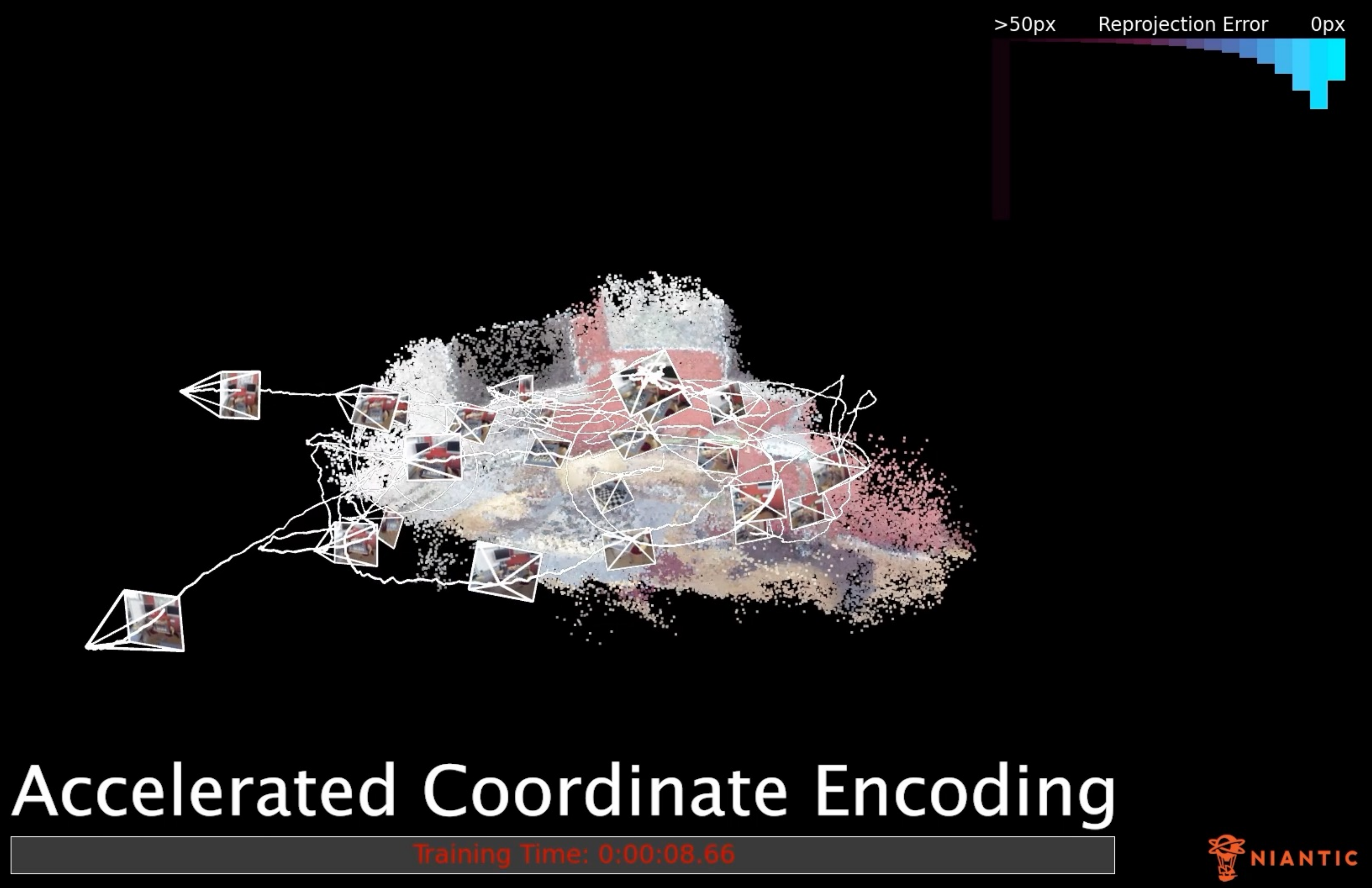


Training Buffer

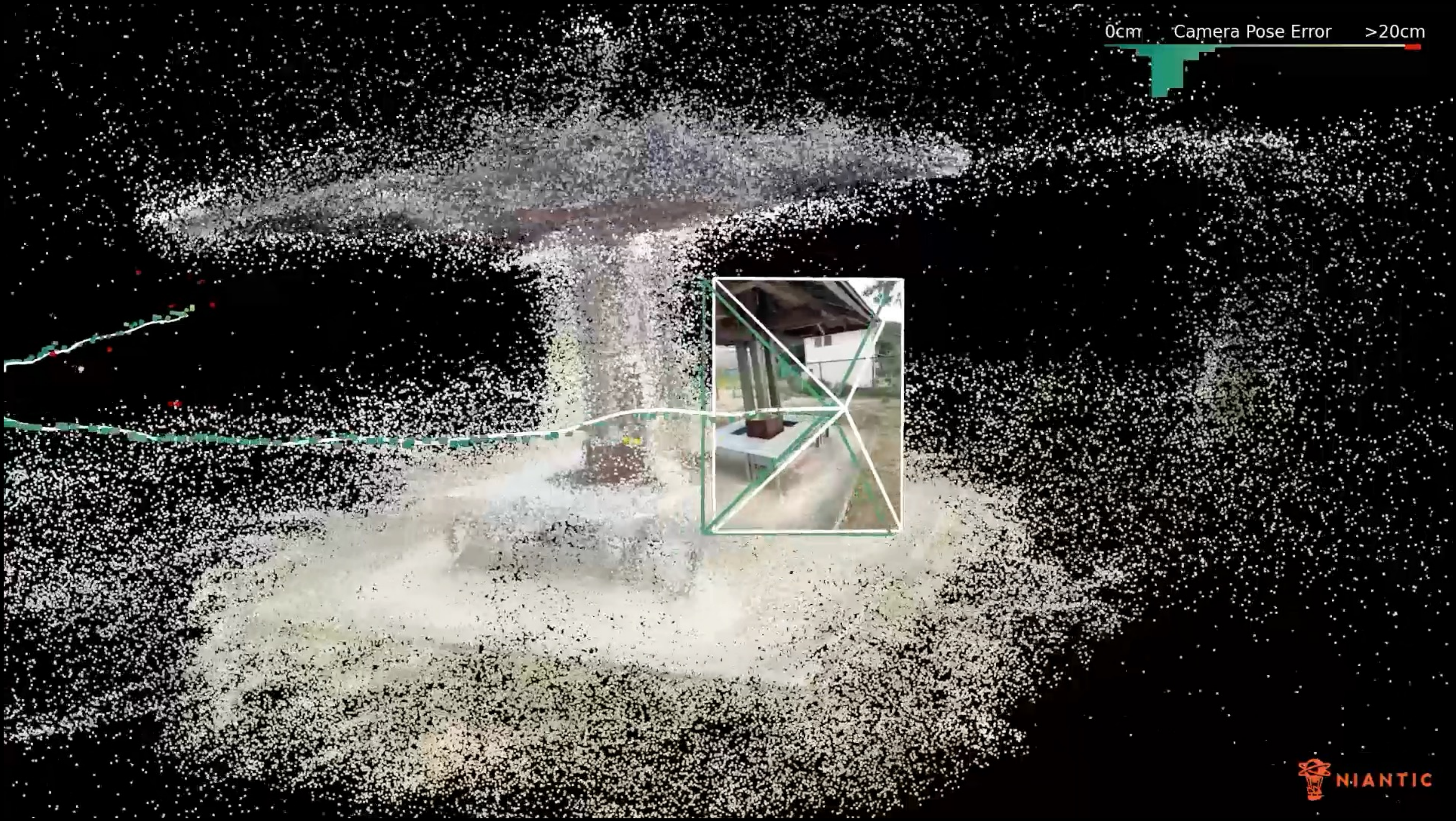
Accelerated Coordinate Encoding



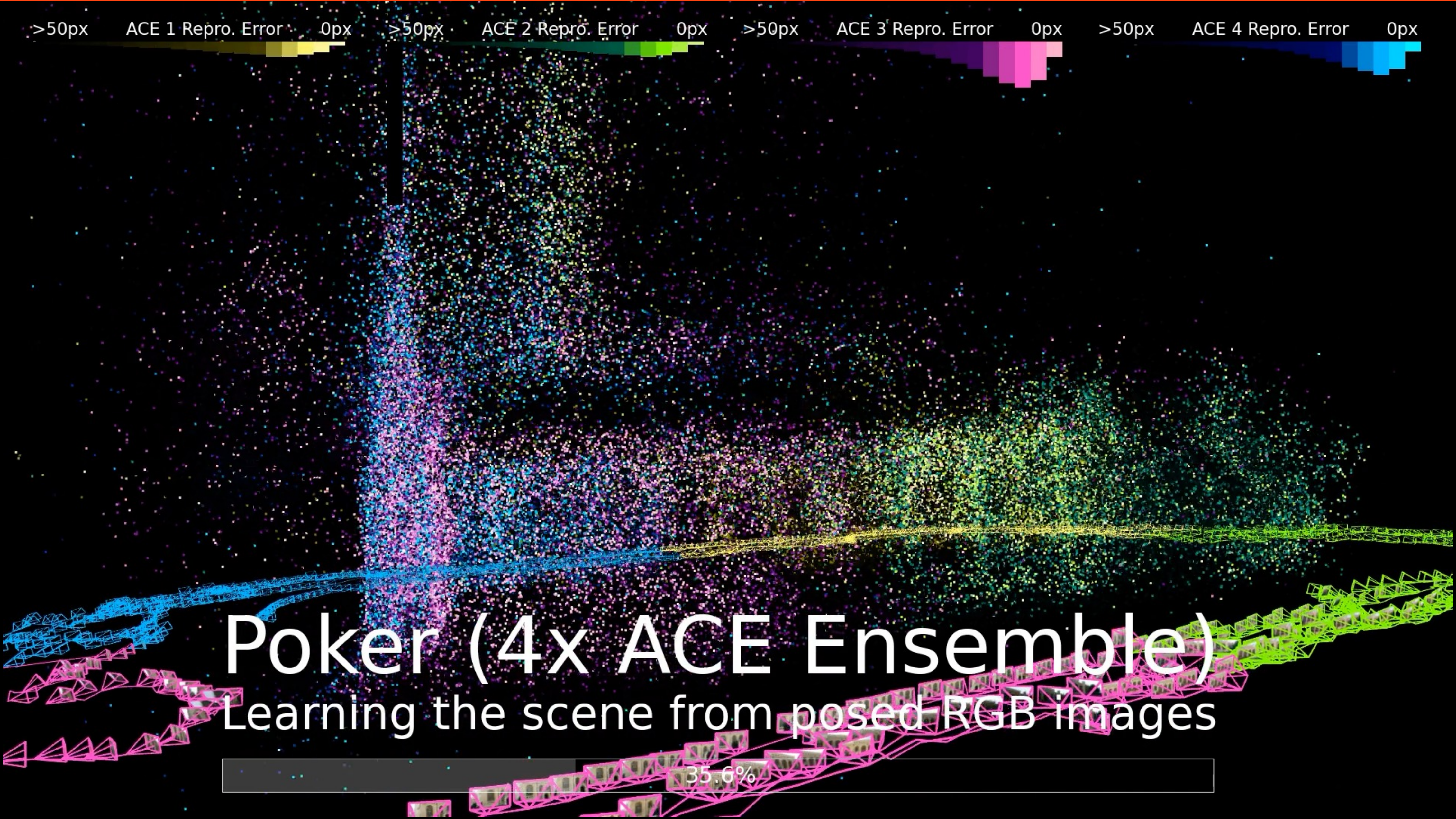


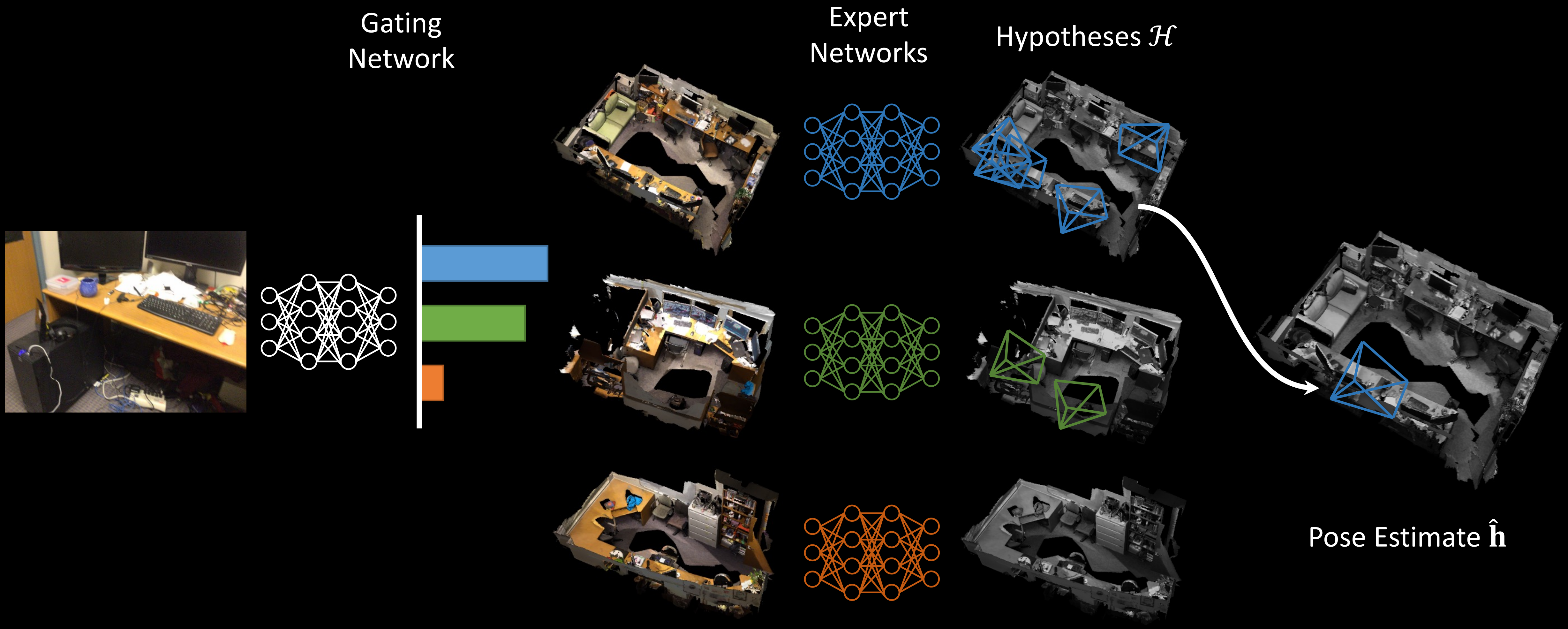


>50px Reprojection Error 0px
Accelerated Coordinate Encoding
Training Time: 0:00:08.66
NIANTIC


0cm
Camera Pose Error
>20cm
NIANTIC

	Mapping w/ Mesh/Depth	Mapping Time	Map Size	7 Scenes		12 Scenes	
				SfM poses	D-SLAM poses	SfM poses	D-SLAM poses
AS (SIFT)	No	~1.5h	~200MB	98.5%	68.7%	99.8%	99.6%
D.VLAD+R2D2	No		~1GB	95.7%	77.6%	99.9%	99.7%
hLoc (SP+SG)	No		~2GB	95.7%	76.8%	100%	99.8%
pixLoc	No		~1GB	N/A	75.7%	N/A	N/A
DSAC* (Full)	Yes	15h	28MB	98.2%	84.0%	99.8%	99.2%
DSAC* (Tiny)	Yes	11h	4MB	85.6%	70.0%	84.4%	83.1%
SANet	Yes	~2.3min	~550MB	N/A	68.2%	N/A	N/A
SRC	Yes	2min‡	40MB	81.1%	55.2%	N/A	N/A
DSAC* (Full)	No	15h	28MB	96.0%	81.1%	99.6%	98.8%
DSAC* (Tiny)	No	11h	4MB	84.3%	69.1%	81.9%	81.6%
ACE	No	5min	4MB	**97.1%**	80.8%	99.9%	99.6%

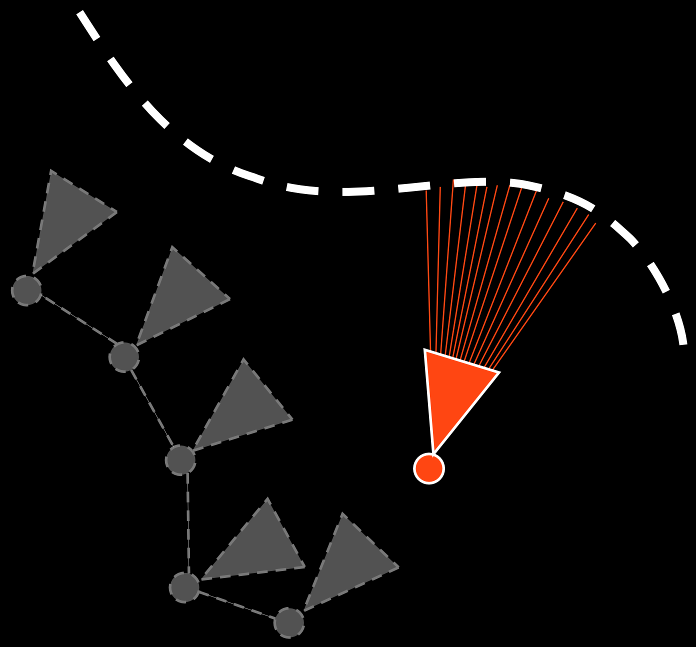

>50px ACE 1 Repro. Error 0px
>50px ACE 2 Repro. Error 0px
>50px ACE 3 Repro. Error 0px
>50px ACE 4 Repro. Error 0px
Poker (4x ACE Ensemble)
Learning the scene from posed RGB images
35.6%



„Expert Sample Consensus Applied to Camera Re-Localization", Brachmann and Rother, ICCV'19

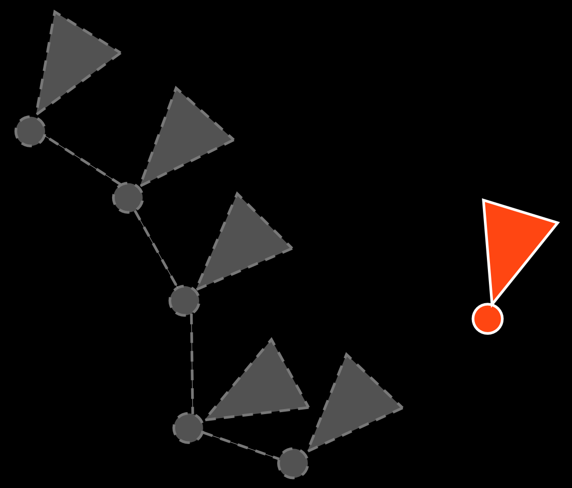
		Mapping w/ Mesh/Depth	Mapping Time	Map Size	Cambridge Landmarks					Avg (cm/°)
					Court	King's	Hospital	Shop	St. Mary's	
FM	AS (SIFT)	No	~35min	~200MB	24/0.1	13/0.2	13/0.2	4/0.2	8/0.3	14/0.2
FM	hLoc (SP+SG)	No	~35min	~800MB	16/0.1	12/0.2	12/0.2	4/0.2	7/0.2	11/0.2
FM	pixLoc	No	~35min	~600MB	30/0.1	14/0.2	14/0.2	5/0.2	10/0.3	15/0.2
FM	GoMatch	No	~35min	~12MB	N/A	25/0.6	25/0.6	48/4.8	335/9.9	N/A
FM	HybridSC	No	~35min	~1MB	N/A	81/0.6	81/0.6	19/0.5	50/0.5	N/A
APR	PoseNet17	No	4-24h	50MB	683/3.5	88/1.0	88/0.1	88/3.8	157/3.3	267/3.0
APR	MS-Transformer	No	~7h	~18MB	N/A	83/1.5	83/1.5	86/3.1	162/4.0	N/A
SCR (w/ Depth)	DSAC* (Full)	Yes	15h	28MB	49/0.3	15/0.3	15/0.3	5/0.3	13/0.4	21/0.3
SCR (w/ Depth)	SANet	Yes	~1min	~260MB	328/2.0	32/0.5	32/0.5	10/0.5	16/0.6	84/0.8
SCR (w/ Depth)	SRC	Yes	2min‡	40MB	81/0.5	39/0.7	39/0.7	19/1.0	31/1.0	42/0.7
SCR	DSAC* (Full)	No	15h	28MB	34/0.2	18/0.3	18/0.3	5/0.3	15/0.6	19/0.4
SCR	DSAC* (Tiny)	No	11h	4MB	98/0.5	27/0.4	27/0.4	11/0.5	56/1.8	45/0.8
SCR	**ACE**	No	5min	4MB	43/0.2	28/0.4	28/0.4	5/0.3	18/0.6	25/0.4
SCR	**Poker (Quad ACE Ensemble)**	No	20min	16MB	28/0.1	18/0.3	18/0.3	5/0.3	9/0.3	17/0.3

Scene Coordinate Regression



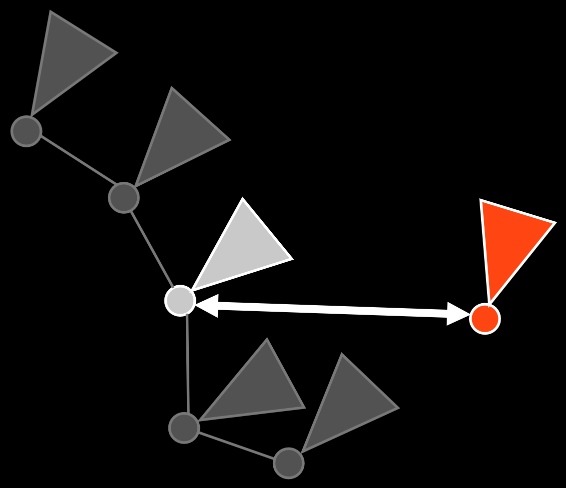
- High Accuracy
- Low Memory Demand
- Fast Mapping
- Large Scale by $n\times$ Small Scale

Absolute Pose Regression



- Can be fast at query time
- Slow at mapping time
- Moderate accuracy (powered by NerF)

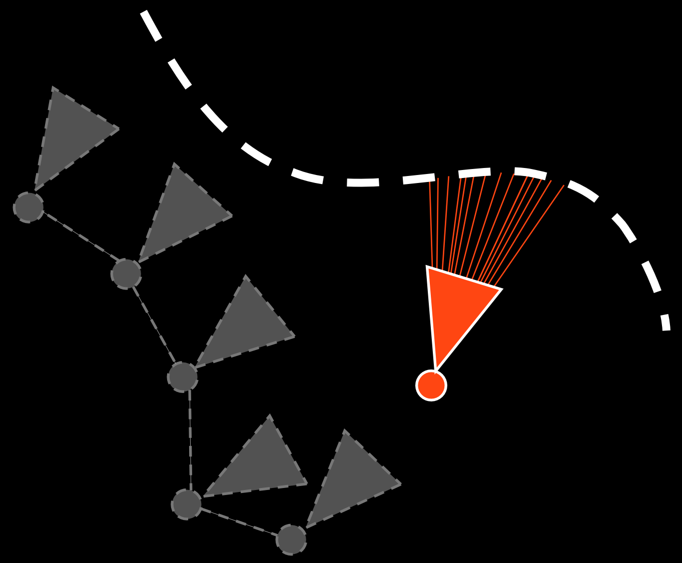
Relative Pose Regression



- Low or no mapping costs
- Scale ambiguity is key challenge
- New dataset and benchmark to measure and drive progress:

  Map-Free Relocalisation

Correspondence Regression (aka Scene Coordinate Regression)



- High Accuracy
- Low Memory Demand
- Fast Mapping
- Large Scale by $n\times$ Small Scale