Part III:
Learning-based Visual Localization

Eric Brachmann

Eric Brachmann

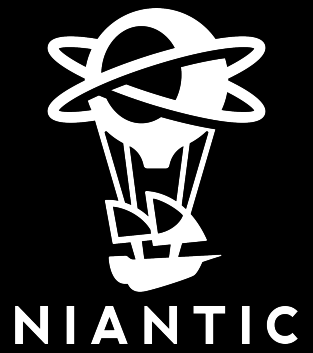
Senior Research Scientist

@eric_brachmann

ebrachmann@nianticlabs.com

https://www.linkedin.com/in/eric-brachmann/

NIANTIC



2001
Keyhole founded

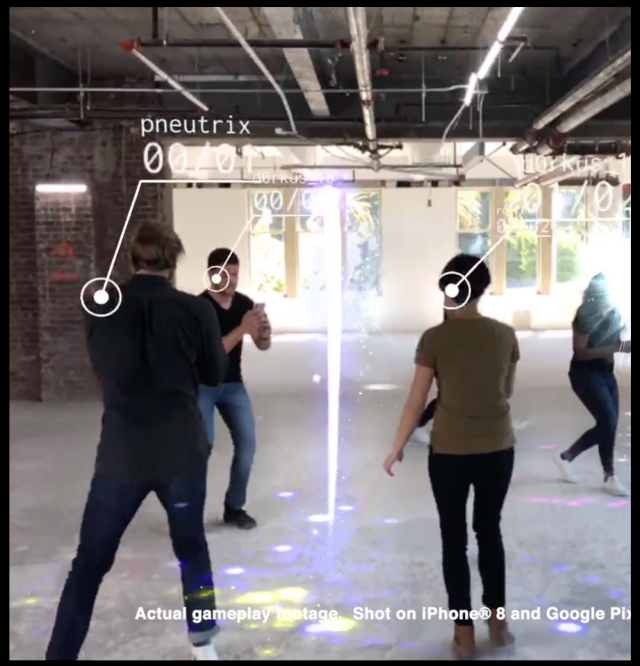
2004
Keyhole acquired (becomes Google Earth)

2011
Niantic incubated at Google

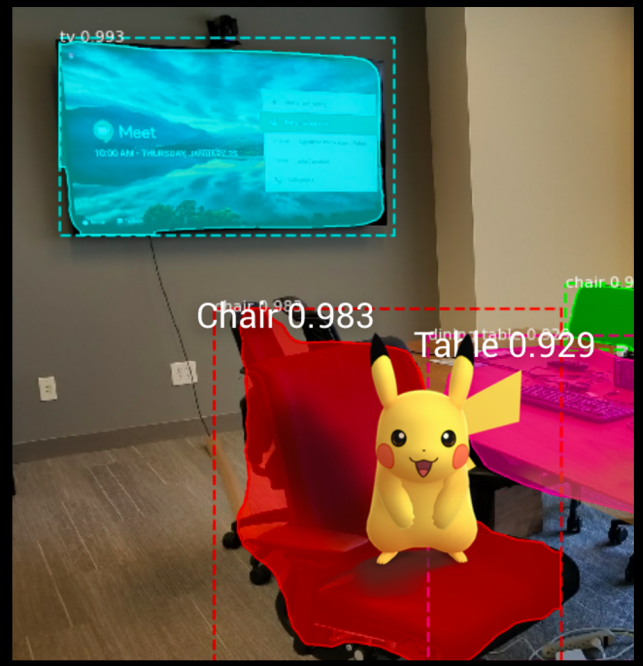
2015
Niantic spins out of Google

2021
Niantic launches Lightship Platform

SHARING
AR EXPERIENCES WITH EACH OTHER



UNDERSTANDING
REALITY TO BETTER AUGMENT IT



MAPPING
TO TIE DIGITAL CONTENT TO REAL WORLD LOCATIONS

Opportunities at https://careers.nianticlabs.com/

Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision
Mehmet Özgur Turkoglu, Eric Brachmann, Konrad Schindler, Gabriel Brostow, Aron Monszpart
3DV 2021

Single Image Depth Prediction with Wavelet Decomposition
Michael Ramamonjisoa, Michael Firman, Jamie Watson, Vincent Lepetit, Daniyar Turmukhambetov
CVPR 2021

Panoptic Segmenta
Colin Graber, Grace
CVPR 2021

Learning to Predict I
Anh-Dzung Doan, D
ICRA 2021

Predicting Visual Ov gs
Anita Rau, Guillermo rmukhambetov
ECCV 2020

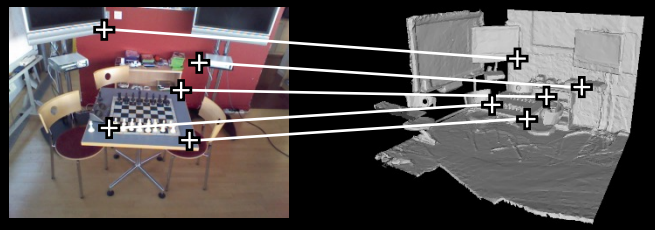
Apply for Internship Program 2022!

More info in November.

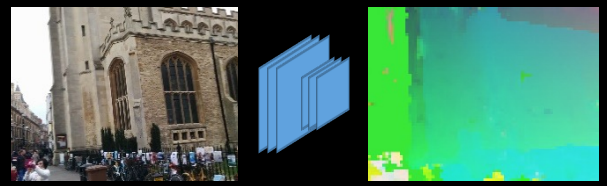
Single-Image Depth Prediction Makes Feature Matching Easier
Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, Fredrik Kahl, Gabriel Brostow
ECCV 2020

Image Stylization for Robust Features
Iaroslav Melekhov, Gabriel Brostow, Juho Kannala, Daniyar Turmukhambetov
ECCV 2020 Workshops

Matching of Hand-Crafted Features
eg. Active Search

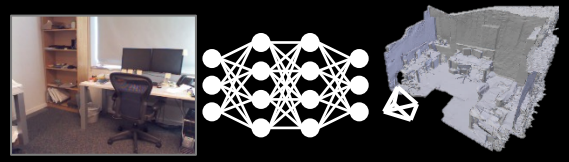


Scene Coordinate Regression
eg. DSAC*



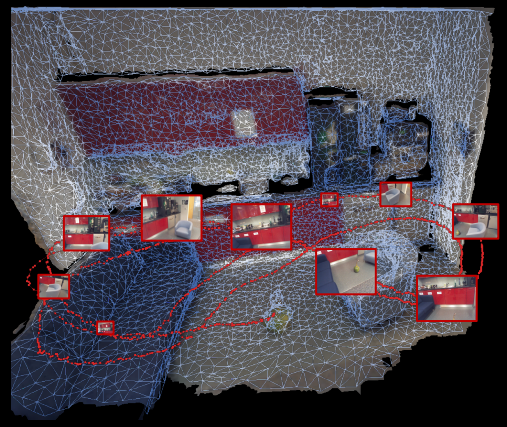
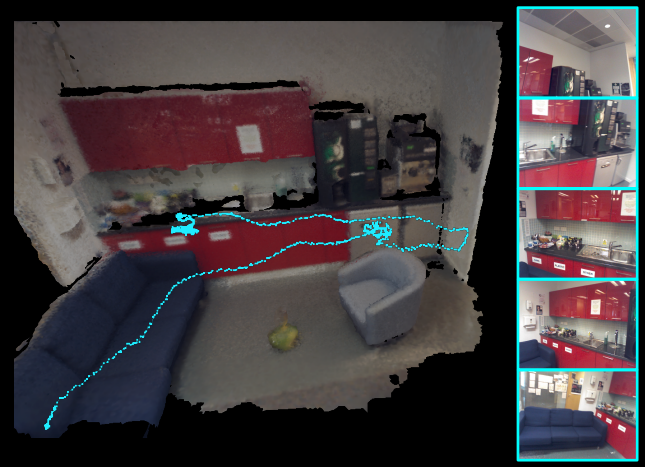
Matching of Learned Features
eg. D2-Net

Absolute Pose Regression
eg. PoseNet



Learn Nothing — Learn Something — Learn Everything

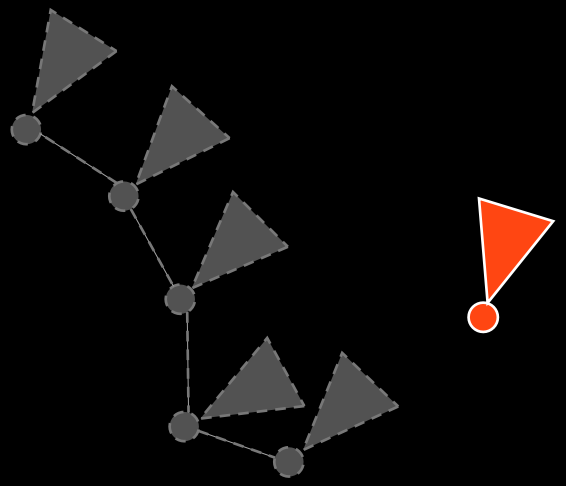
[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[D2-Net] “D2-Net: A Trainable CNN for Joint Description and Detection of Local Features”, Dusmanu et al., CVPR’19
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17

	Preparation Scene-Agnostic Training	Mapping Scene-Specific Training	Re-Localisation Evaluation
			
Active Search	Build descriptor dictionary	Triangulate Scene	Discrete Feature Matching
hLoc	Train SuperPoint Train SuperGlue Train NetVLAD	Build Retrieval Index Triangulate Scene	NN Retrieval Discrete Feature Matching

[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[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

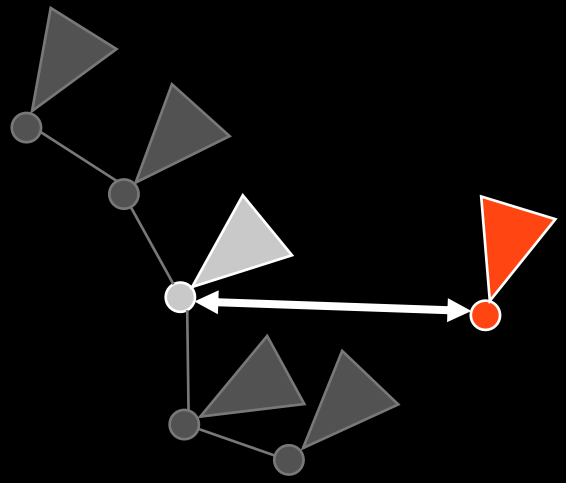
Absolute Pose Regression

The challenge.



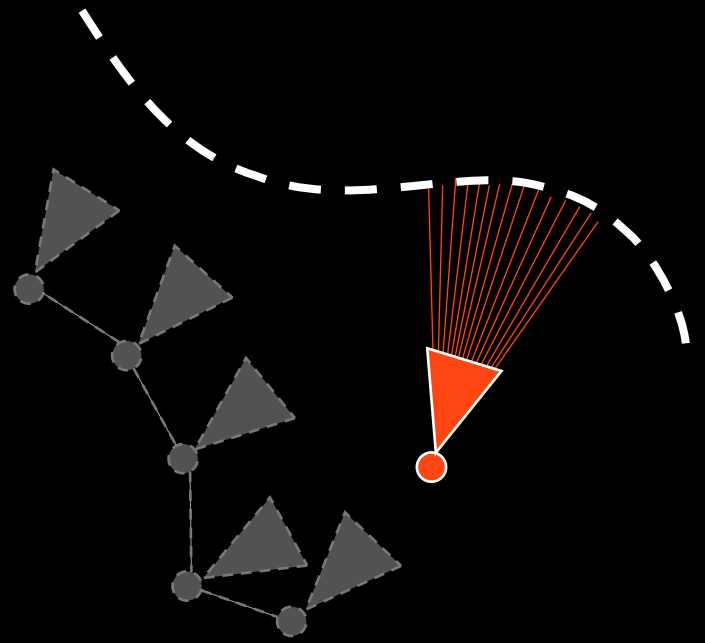
Relative Pose Regression

The promise.



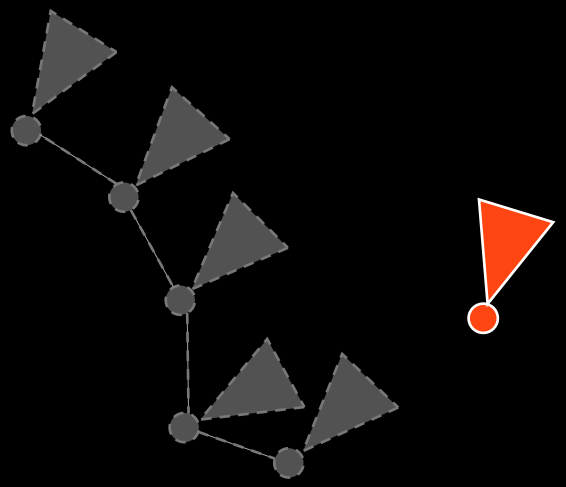
Scene Coordinate Regression

The compromise.



Absolute Pose Regression

The challenge.

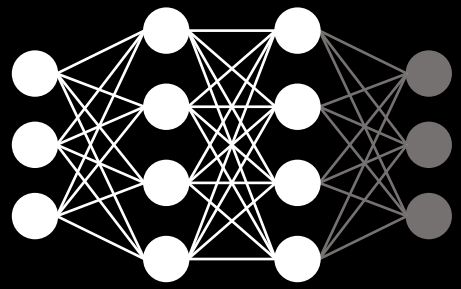


- What is absolute pose regression?
- Tool box
  - Pose parametrization
  - Pose loss
- What can we achieve?
- The challenge

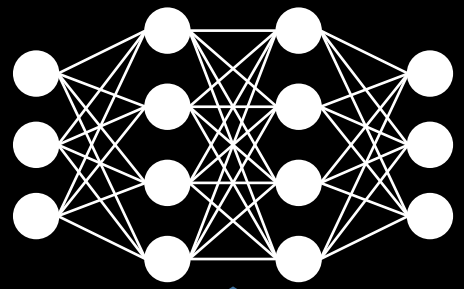
Preparation	Mapping	Re-Localisation
Scene-Agnostic Training	Scene-Specific Training	Evaluation

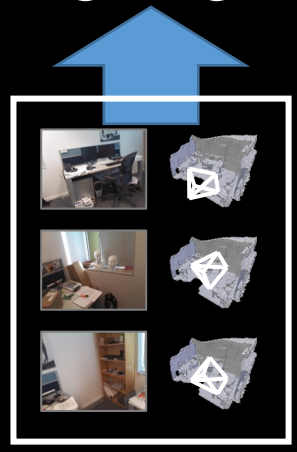
PoseNet

Pre-Train Backbone



Train Pose Regression





Training Data

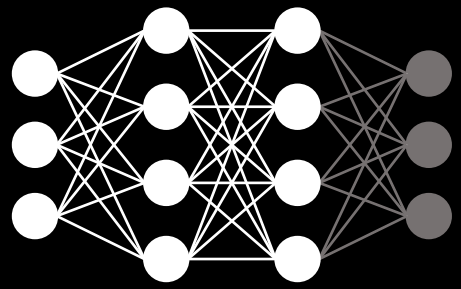
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17

Preparation
Scene-Agnostic Training

Mapping
Scene-Specific Training

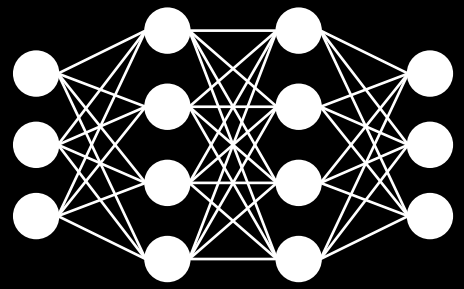
Re-Localisation
Evaluation

Pre-Train Backbone

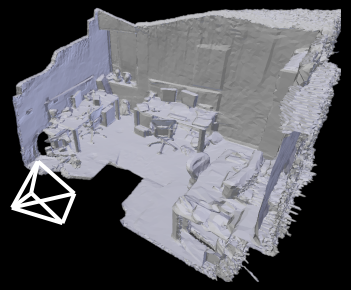


Train Pose Regression

Input:
Image I



Forward Pass



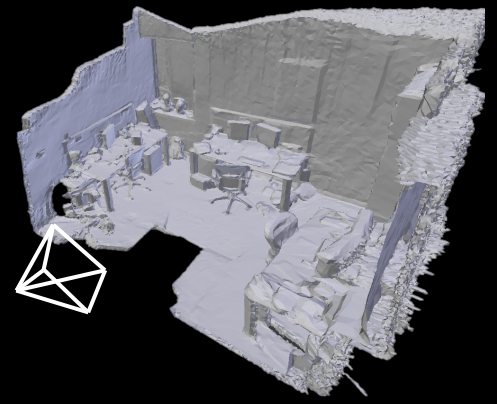
Output:
Pose $\hat{\mathbf{h}}$

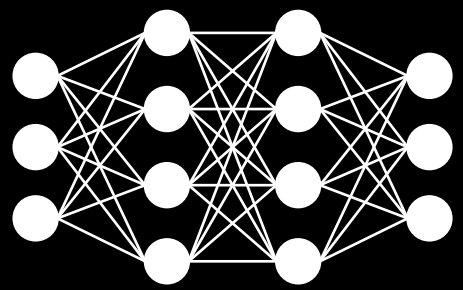
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17


The Toolbox

$$\mathbf{h} \in \mathrm{SE}(3) \text{ with } \mathrm{SE}(3) = \left\{ \begin{pmatrix} R & \mathbf{t} \\ 0 & 1 \end{pmatrix} | R \in SO(3), \mathbf{t} \in \mathbb{R}^3 \right\}$$

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





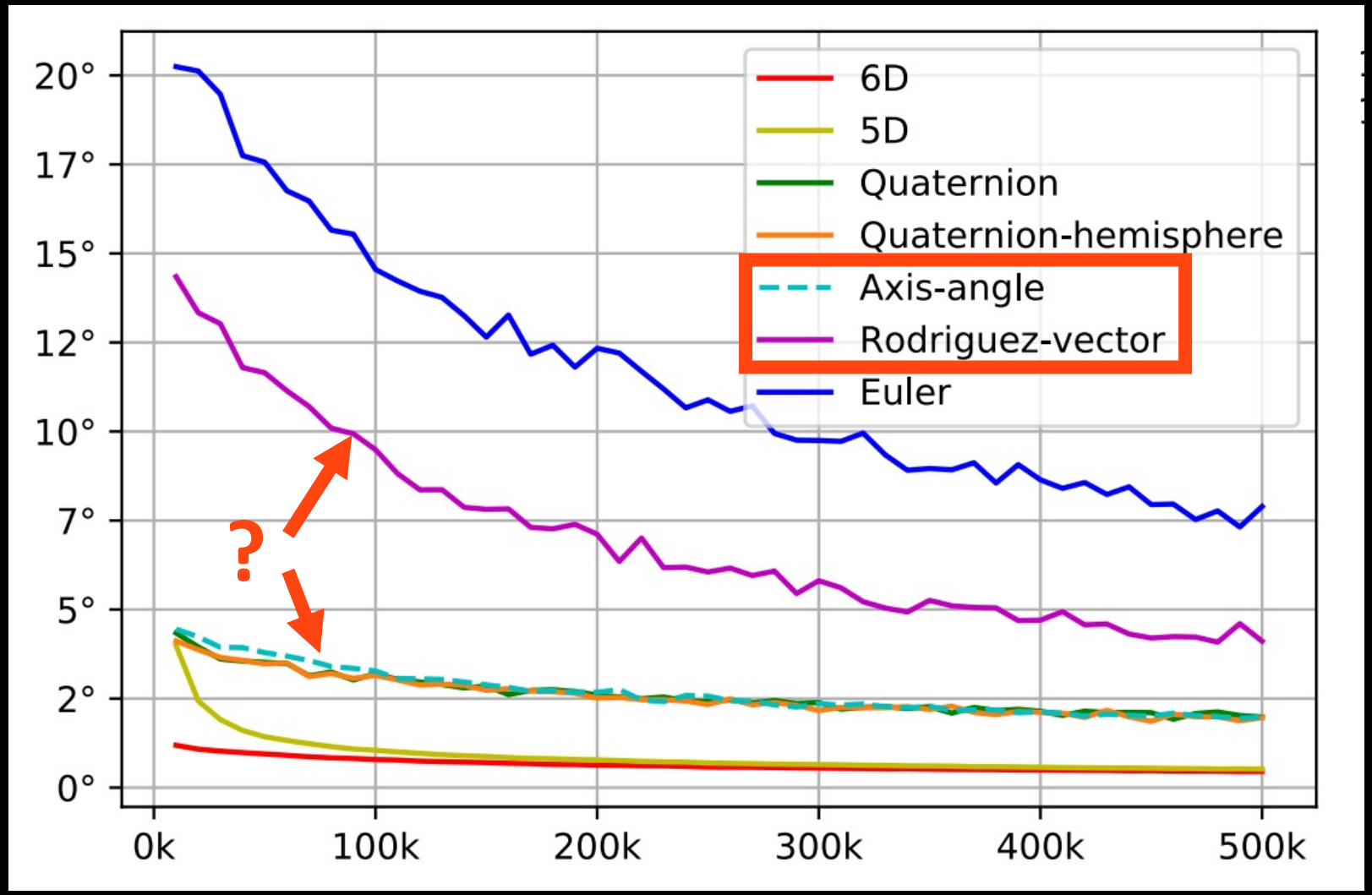
Rotation Matrix	Unit Quaternion e.g. [1]	Axis-Angle e.g. [2]	Log Unit Quaternion e.g. [3]
$R = \begin{pmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{pmatrix}$	$\mathbf{q} = (c, v_1, v_2, v_3)$	$\log R = \theta\hat{\mathbf{u}} = (u_1', u_2', u_3')$	$\log \mathbf{q} = 2 \log R$ [4]
Enforce/Map: $RR^{\mathrm{T}} = I, \det(\mathrm{R}) = 1$	Enforce/Map: $\|\mathbf{q}\| = 1$	Enforce/Map: –	Enforce/Map: –

6D (e.g. in [5]) → (partial) Gram-Schmidt
9D (e.g. in [6]) → orthogonal Procrustes

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

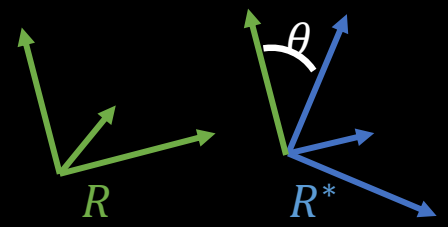
[1] Kendall et al., "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization", ICCV15
[2] Brachmann et al., "DSAC - Differentiable RANSAC for Camera Localization", CVPR17
[3] Brahmbhatt et al., "Geometry-Aware Learning of Maps for Camera Localization", CVPR18
[5] Zhou et al., "On the Continuity of Rotation Representations in Neural Networks", CVPR'19
[6] Chen et al., "Wide-Baseline Relative Camera Pose Estimation with Directional Learning", CVPR'21

MLP trained to map SO(3) to X



Zhou et al., “On the Continuity of Rotation Representations in Neural Networks”, CVPR’19

$\mathbf{h} = (R, \mathbf{t}) \quad \ell(\mathbf{t}, \mathbf{t}^*) = \|\mathbf{t} - \mathbf{t}^*\|$

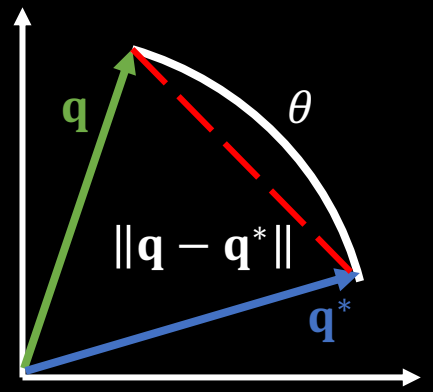


How to measure rotation error?

Angular Distance [1]: $\theta(R, R^*) = \left\|\log(R^* R^{\mathrm{T}})\right\|$

Quaternion Distance [2]: $\|\mathbf{q} - \mathbf{q}^*\|$

Angle-Axis Distance,
Log Quaternion Distance [3]: $\|\log R - \log R^*\|$



$$\|\log R - \log R^*\| \neq \|\log TR - \log TR^*\| \text{ [4]}$$

[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] Brahmbhatt et al., “Geometry-Aware Learning of Maps for Camera Localization”, CVPR18
[4] Hartley et al., „Rotation Averaging“, IJCV13

How to combine rotation error and translation error?

Hand-Tuned [1]:

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

Self-Tuned [2]:

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

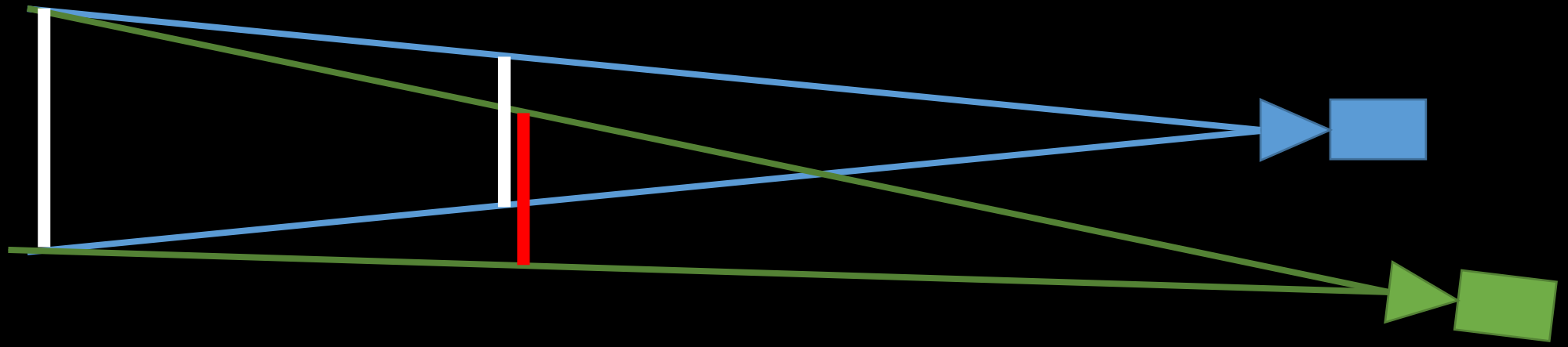
$$s = \log \sigma^2$$

[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", CVPR 2017

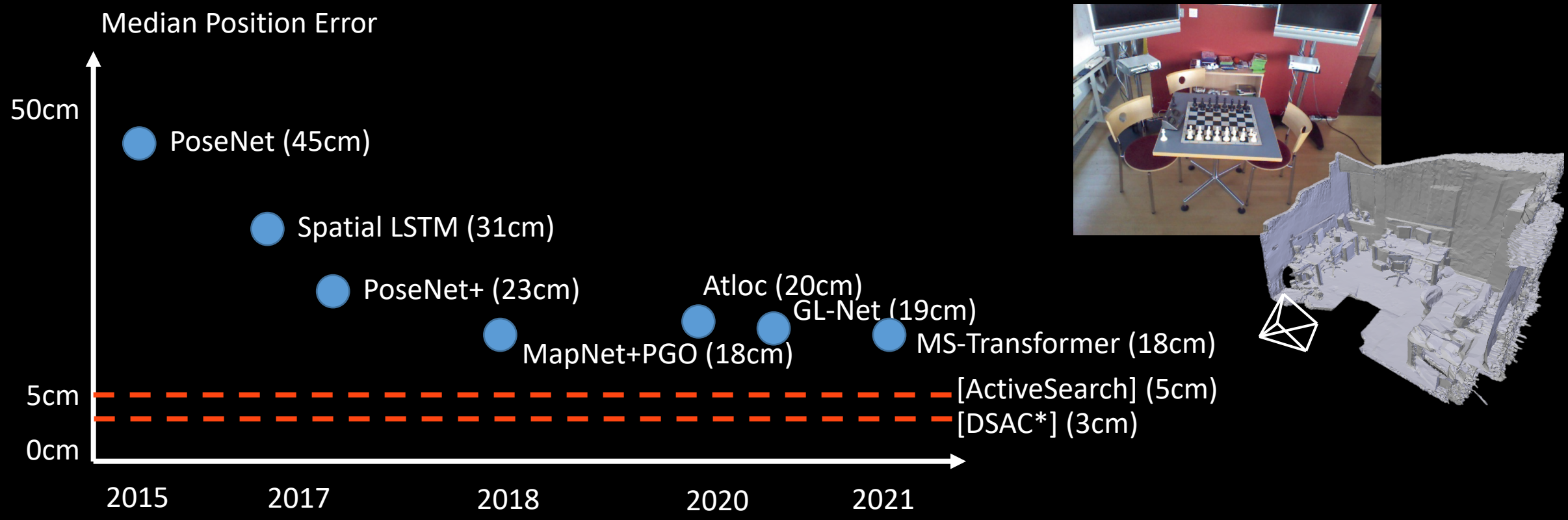
How to combine rotation error and translation error?

Measure the reprojection error [1]:

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



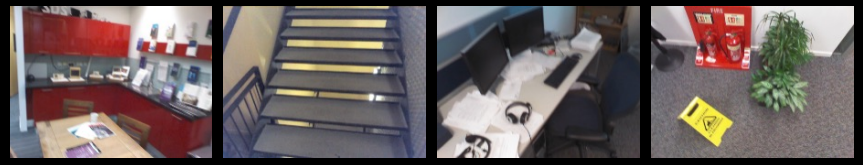
[1] Kendall and Cipolla, “Geometric Loss Functions for Camera Pose Regression with Deep Learning”, CVPR 2017

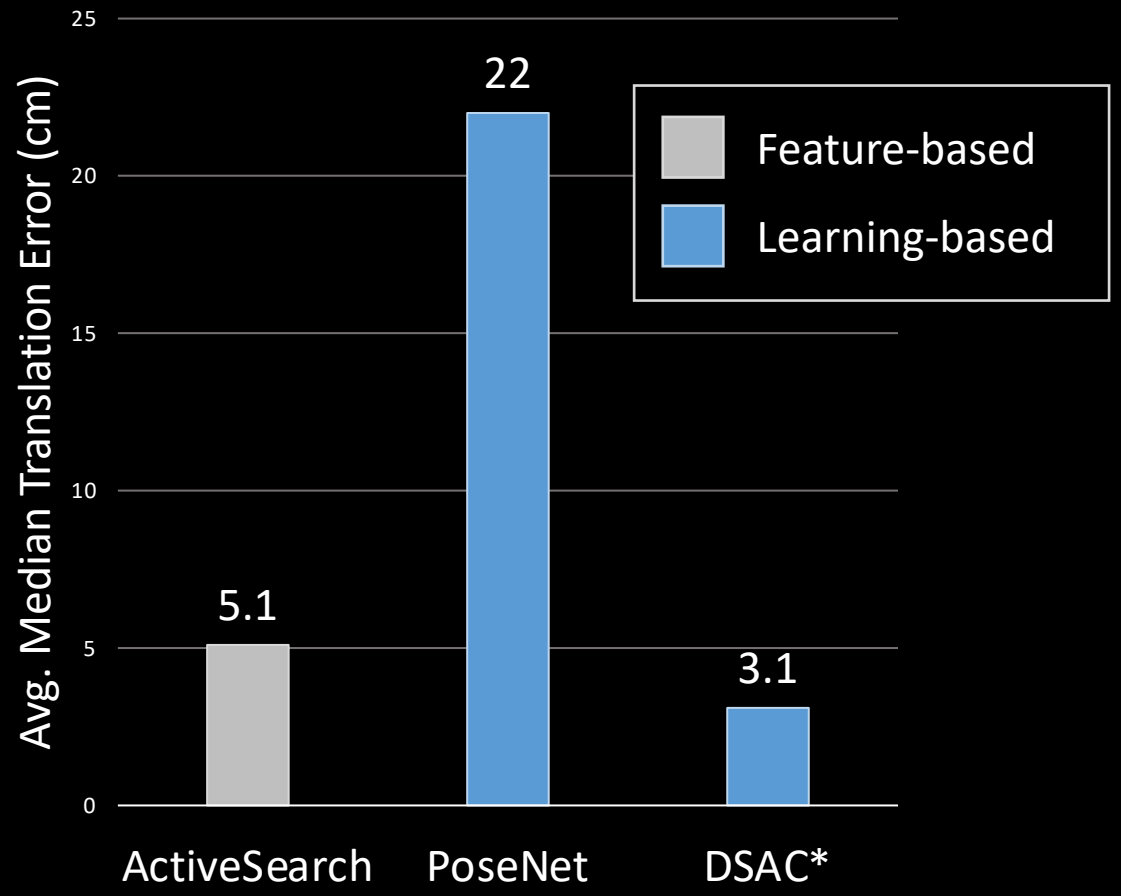


[PoseNet]"PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Localization", Kendall et al., ICCV 2015
[Spatial LSTM]"Image-Based Localization with Spatial LSTMs", Walch et al., ICCV 2017
[PoseNet+]"Geometric Loss Functions for Camera Pose Regression with Deep Learning", Kendall and Cipolla, CVPR 2017
[MapNet]"Geometry-Aware Learning of Maps for Camera Localization", Brahmbhatt et al., CVPR18
[Atloc]"Atloc: Atten-tion guided camera localization", Wang et al., AAAI 2020
[AbsPoseGNN] "Learn-ing multi-view camera relocalization with graph neural networks", Xue et al., CVPR 2020
[MS-Transformer] "Learning Multi-Scene Absolute Pose Regression with Transformers", Shavit et al., ICCV 2021

[ActiveSearch]"Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization", Sattler et al., PAMI 2017
[DSAC*] "Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC", Brachmann and Rother, TPAMI'21

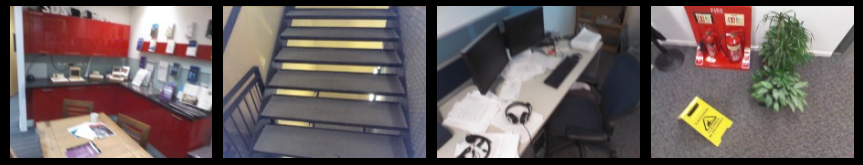
7Scenes Dataset [Sho13]

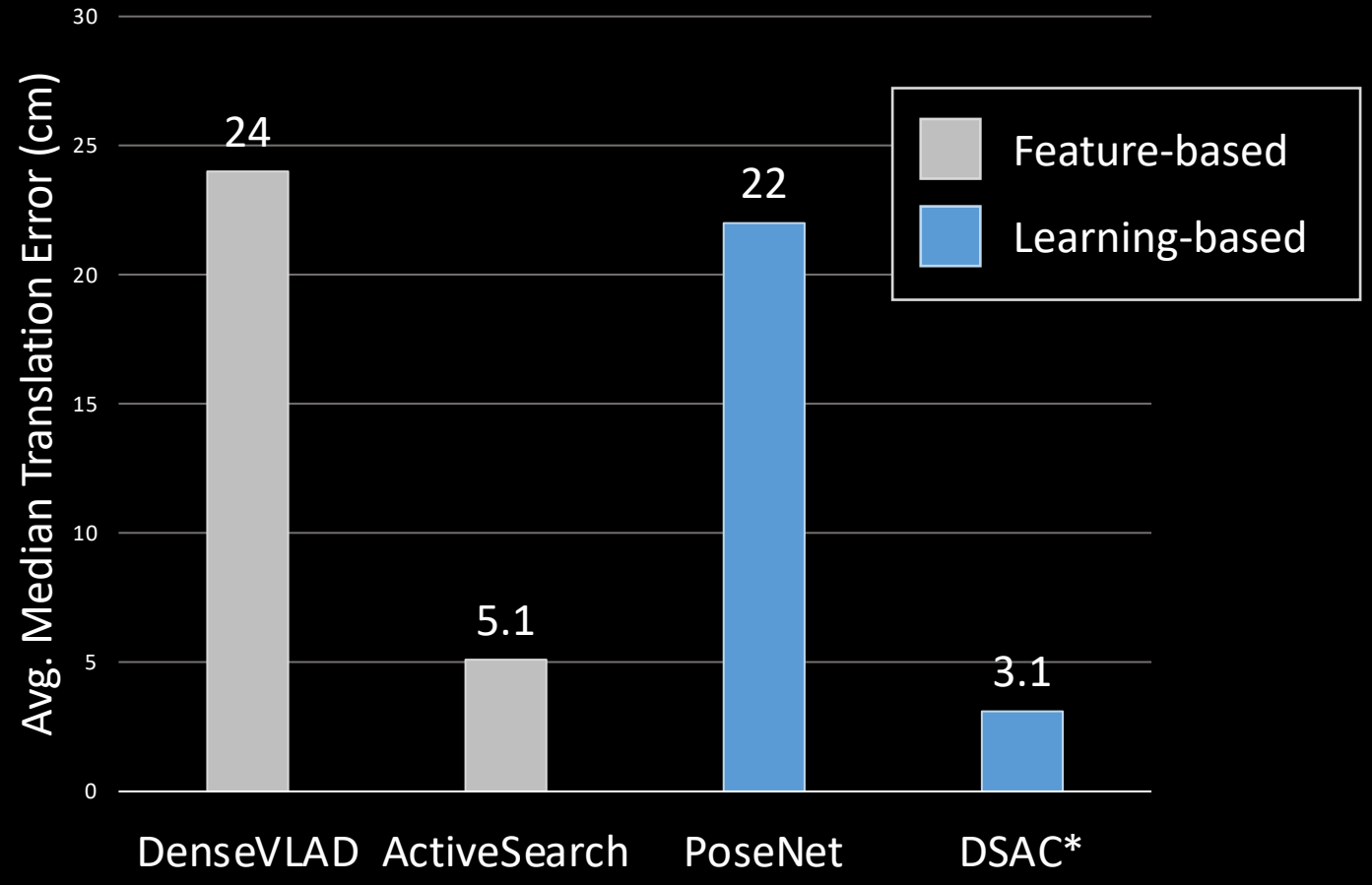




[Sho13] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17
[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21
[DenseVLAD] “24/7 Place Recognition by View Synthesis”, Torii et al., TPAMI’18

7Scenes Dataset [Sho13]

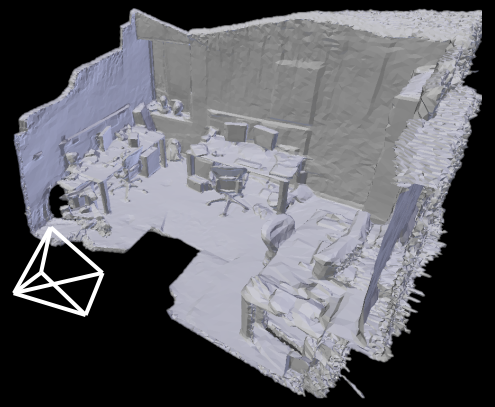




[Sho13] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17
[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21
[DenseVLAD] “24/7 Place Recognition by View Synthesis”, Torii et al., TPAMI’18

Layers $L_0 - L_n$

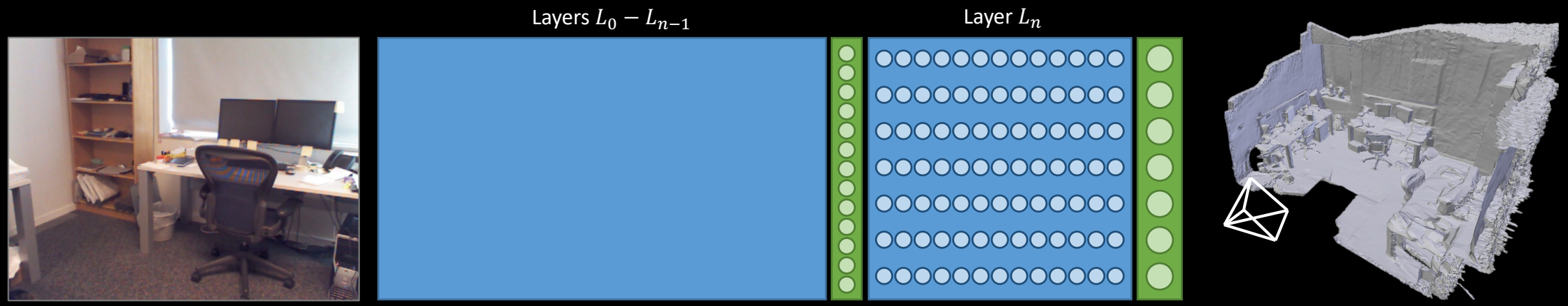
Absolute Pose Regression Network



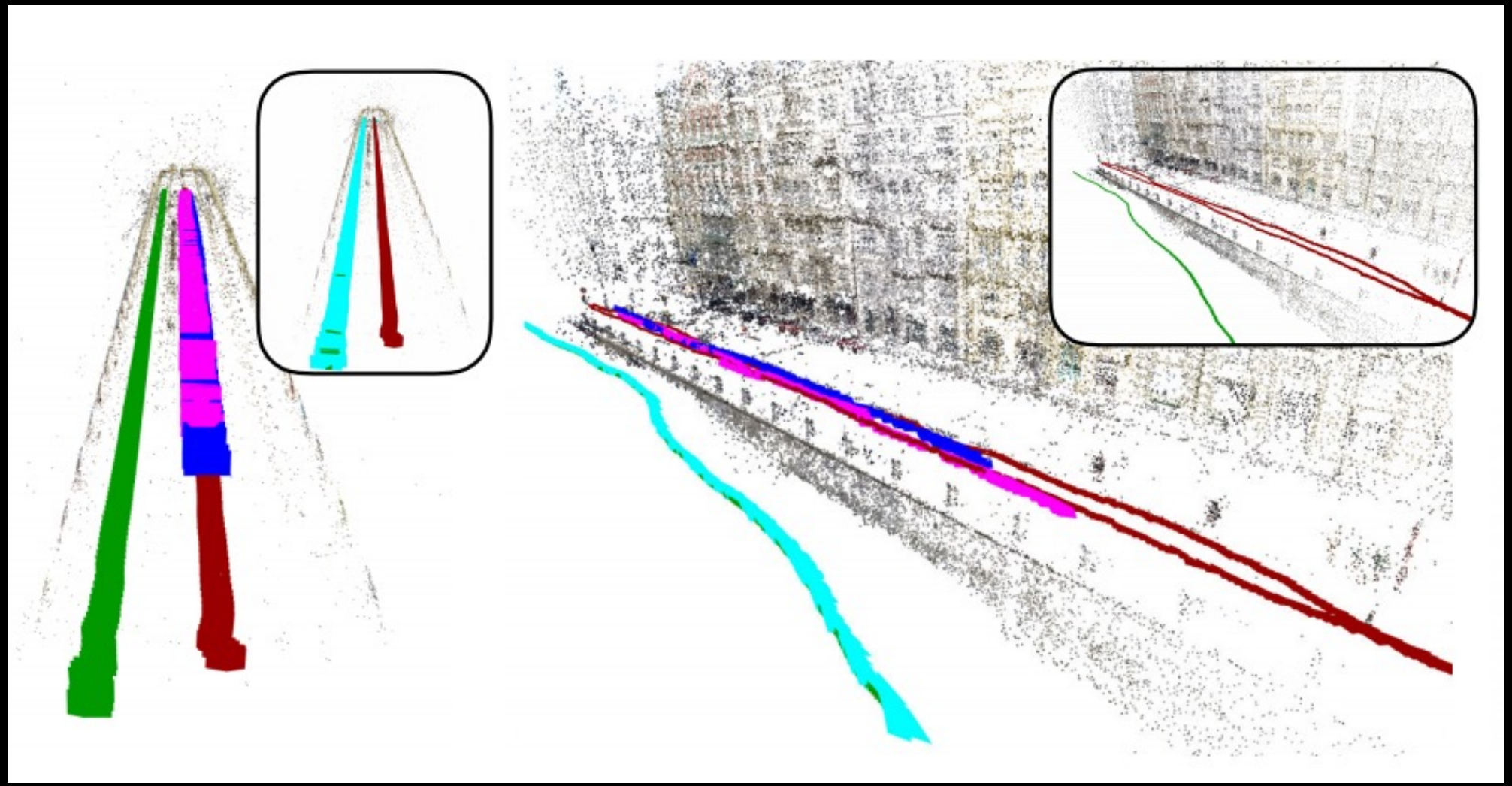
Input: Image I

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

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

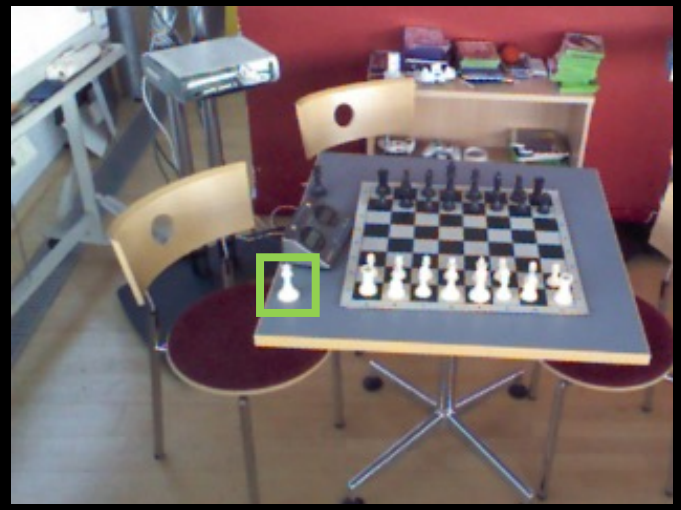


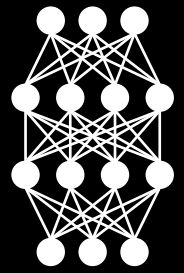
“Understanding the Limitations of CNN-based Absolute Camera Pose Regression”, Sattler et al., CVPR’19



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

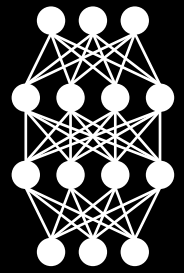
Training Image



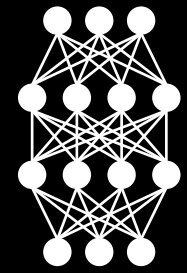


$\hat{\mathbf{h}}$ 

Test Images



???



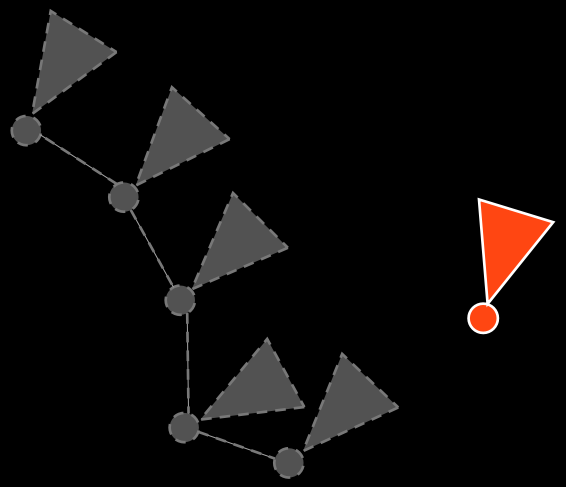
???

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

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

Absolute Pose Regression

The challenge.



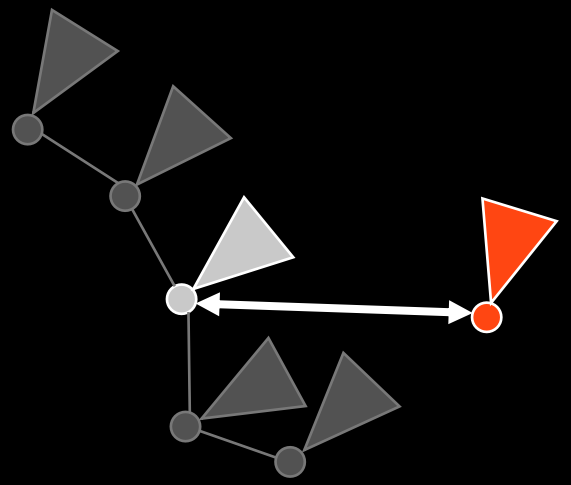
- Fast at query time (few ms)
- Low accuracy
- Insufficient training data?

# training images	203	683	2,035	7,425
PoseNet [29]	1.19 / 6.88	1.15 / 8.10	0.86 / 6.88	0.54 / 5.84
MapNet [11]	1.07 / 4.70	0.72 / 3.41	0.42 / 2.06	0.38 / 2.31
Active Search [59]	**0.01 / 0.04**			
DenseVLAD [71]	0.98 / 7.90	0.72 / 7.81	0.61 / 7.38	0.57 / 6.94
DenseVLAD+Inter.	0.89 / 5.71	0.57 / 5.96	0.48 / 6.13	0.41 / 6.41

Synthetic experiment from “Understanding the Limitations of CNN-based Absolute Camera Pose Regression”, Sattler et al., CVPR’19

Relative Pose Regression

The promise.



- What is relative pose regression?
- Where do we stand?
- A traditional baseline

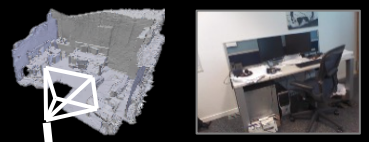
RelocNet

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

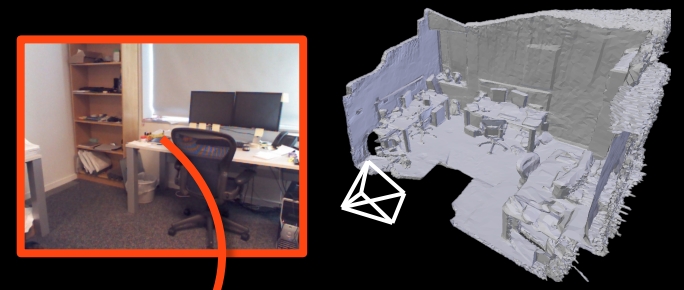
Pre-Train Image Retrieval

Pre-Train Relative Pose Regressor

Build Retrieval Index



Retrieve NN
Refine pose



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

		# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
	DSAC* [10]*	1	0.02, 1.1°	0.02, 1.2°	0.01, 1.8°	0.03, 1.2°	0.04, 1.4°	0.03, 1.7°	0.04, 1.4°	0.03, 1.4°
Seq. based	VidLoc [20]*	200	0.18, -	0.26, -	0.14, -	0.26, -	0.36, -	0.31, -	0.26, -	0.25, -
Seq. based	LsG [70]*	7	0.09, 3.3°	0.26, 10.9°	0.17, 12.7°	0.18, 5.5°	0.20, 3.7°			
Seq. based	MapNet [11]*	3	0.08, 3.3°	0.27, 11.7°	0.18, 13.3°	0.17, 5.2°	0.22, 4.0°			
Seq. based	GL-Net [71]*?	8	0.08, 2.8°	0.26, 8.9°	0.17, 11.4°	0.18, 5.1°	0.15, 2.8°			
Image based APR	PoseNet [35]*	1	0.32, 6.6°	0.47, 14.0°	0.30, 12.2°	0.48, 7.2°	0.49, 8.1°			
Image based APR	Bayesian PoseNet [33]*	1	0.37, 7.2°	0.43, 13.7°	0.31, 12.0°	0.48, 8.0°	0.61, 7.1°			
Image based APR	Geometric PoseNet [34]*	1	0.13, 4.5°	0.27, 11.3°	0.17, 13.0°	0.19, 5.6°	0.26, 4.8°			
Image based APR	MLFBPPose [67]*	1	0.12, 5.8°	0.26, 12.0°	0.14, 13.5°	0.18, 8.2°	0.21, 7.1°			
Image based APR	Hourglass [45]*	1	0.15, 6.2°	0.27, 10.8°	0.19, 11.6°	0.21, 8.5°	0.25, 7.0°			
Image based APR	LSTM-Pose [64]*	1	0.24, 5.8°	0.34, 11.9°	0.21, 13.7°	0.30, 8.1°	0.33, 7.0°			
Image based APR	BranchNet [68]*	1	0.18, 5.2°	0.34, 9.0°	0.20, 14.2°	0.30, 7.1°	0.27, 5.1°			
Image based APR	ANNet [12]*	1	0.12, 4.3°	0.27, 11.6°	0.16, 12.4°	0.19, 6.8°	0.21, 5.2°			
Image based APR	GPoseNet [13]*	1	0.20, 7.1°	0.38, 12.3°	0.21, 13.8°	0.28, 8.8°	0.37, 6.9°	0.35, 8.2°	0.37, 12.5°	0.31, 10.0°
Image based APR	AttLoc [65]*	1	0.10, 4.1°	0.25, 11.4°	0.16, 11.8°	0.17, 5.3°	0.21, 4.4°	0.23, 5.4°	0.26, 10.5°	0.20, 7.6°
Image based APR	AnchorPoint [49]*?	1	**0.06**, 3.9°	**0.16**, 11.1°	**0.09**, 11.2°	**0.11**, 5.4°	**0.14**, 3.6°	**0.13**, 5.3°	**0.21**, 11.9°	0.13, 7.5°
IR	DenseVLAD [59]	1	0.21, 12.5°	0.33, 13.8°	0.15, 14.9°	0.28, 11.2°	0.31, 11.2°	0.30, 11.3°	0.25, 12.3°	0.26, 12.5°
IR	DenseVLAD+Inter [55]	1	0.18, 10.0°	0.33, 12.4°	0.14, 14.3°	0.25, 10.1°	0.26, 9.4°	0.27, 11.1°	0.24, 14.7°	0.24, 11.7°
RPR	NN-Net [38]	1	0.13, 6.5°	0.26, 12.7°	0.14, 12.3°	0.21, 7.4°	0.24, 6.4°	0.24, 8.0°	0.27, 11.8°	0.21, 9.3°
RPR	RelocNet [3]	1	0.12, 4.1°	0.26, 10.4°	0.14, 10.5°	0.18, 5.3°	0.26, 4.2°	0.23, 5.1°	0.28, 7.5°	0.21, 6.7°
RPR	EssNet [73]	1	0.13, 5.1°	0.27, 10.1°	0.15, 9.9°	0.21, 6.9°	0.22, 6.1°	0.23, 6.9°	0.32, 11.2°	0.22, 8.0°
RPR	EssNet [73] reprod.	1	-	-	-	-	-	-	0.32, 9.8°	-
RPR	NC-EssNet [73]	1	0.12, 5.6°	0.26, 9.6°	0.14, 10.7°	0.20 , 6.7°	0.22, 5.7°	0.22, 6.3°	0.31, 7.9°	0.21, 7.5°
RPR	NC-EssNet [73] reprod.	1	0.13, 5.5°	-	-	-	-	-	-	-
RPR	CamNet [23]?	1	-	-	-	-	-	-	-	**0.05**, **1.8°**
RPR	RelPose GNN	1	0.08, **2.7°**	0.21, **7.5°**	0.13, **8.7°**	0.15, **4.1°**	0.15, **3.5°**	0.19, **3.7°**	0.22, **6.5°**	0.16, 5.2°

Essential Matrix Estimation	Training Data	Testing Data: Cambridge	Testing Data: 7Scenes
EssNet	Cambridge	1.08/3.41	0.57/80.06
NC-EssNet	Cambridge	0.85/2.82	0.48/32.97
EssNet	7Scenes	10.36/85.75	0.22/8.03
NC-EssNet	7Scenes	7.98/24.35	0.21/7.50

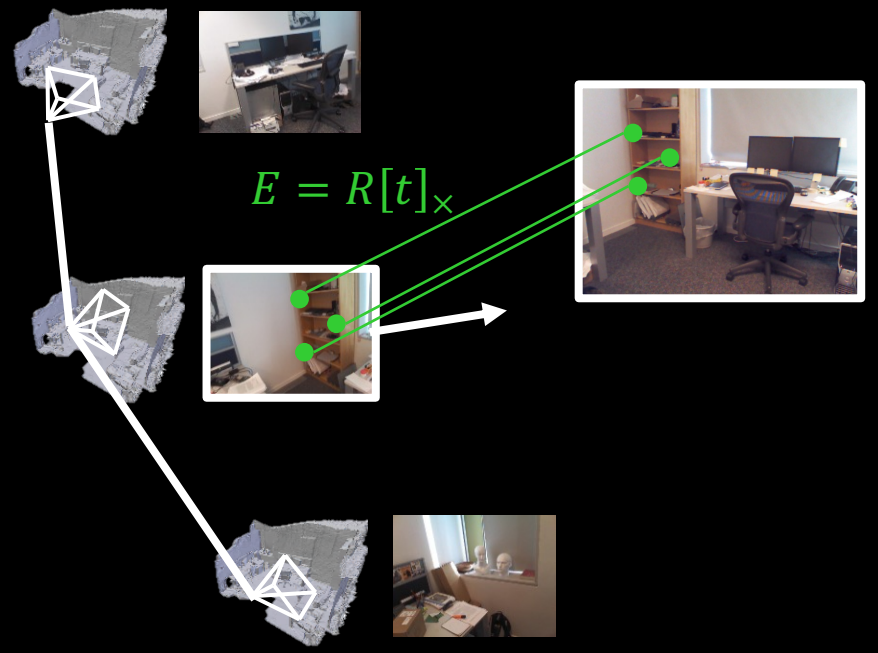
From “To Learn or Not to Learn: Visual Localization from Essential Matrices”, Zhou at al., ICRA’20

RelocNet (trained on ScanNet)
0.29, 11.3°

“Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision”, Turkoglu et al., 3DV’21

Preparation Scene-Agnostic Training	Mapping Scene-Specific Training	Re-Localisation Evaluation
Nothing.	Build Retrieval Index (DenseVLAD)	Retrieve NN Refine pose

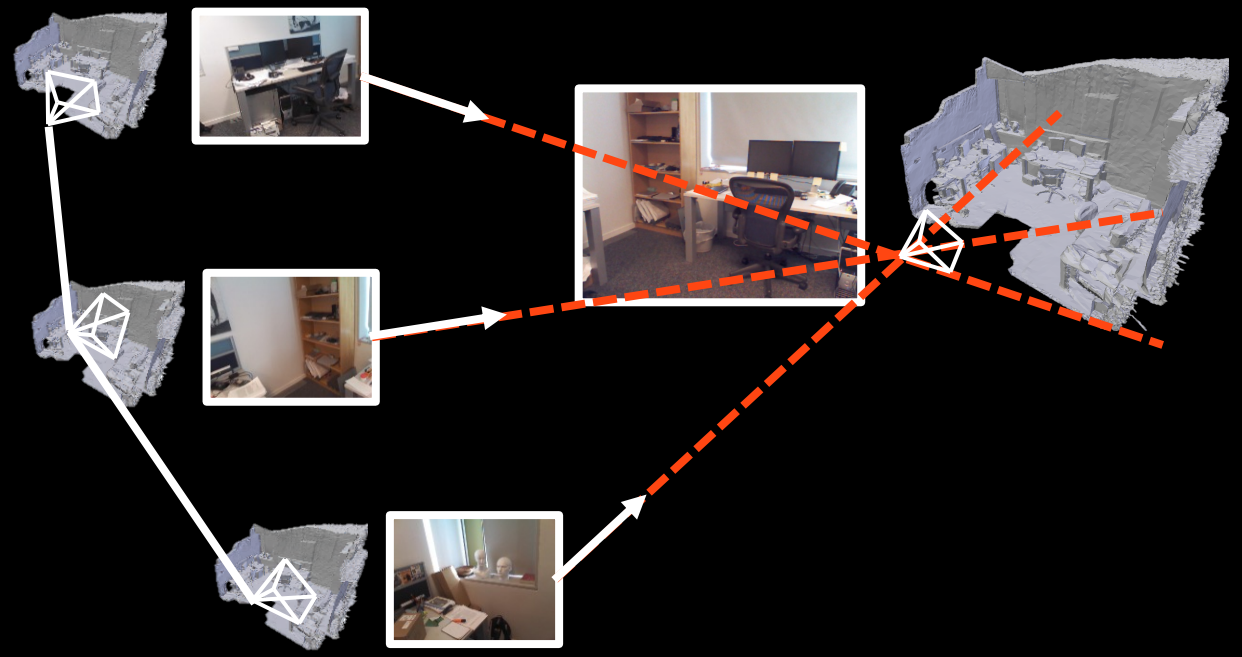
SIFT+5Point



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

Preparation Scene-Agnostic Training	Mapping Scene-Specific Training	Re-Localisation Evaluation
Nothing.	Build Retrieval Index (DenseVLAD)	Retrieve NN Refine pose

SIFT+5Point



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

		# test frames	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs	Avg.
	DSAC* [10]*	1	0.02, 1.1°	0.02, 1.2°	0.01, 1.8°	0.03, 1.2°	0.04, 1.4°	0.03, 1.7°	0.04, 1.4°	0.03, 1.4°
Seq. based	VidLoc [20]*	200	0.18, -	0.26, -	0.14, -	0.26, -	0.36, -	0.31, -	0.26, -	0.25, -
	LsG [70]*	7	0.09, 3.3°	0.26, 10.9°	0.17, 12.7°	0.18, 5.5°	0.20, 3.7°	0.23, 4.9°	0.23, 11.3°	0.19, 7.5°
	MapNet [11]*	3	0.08, 3.3°	0.27, 11.7°	0.18, 13.3°	0.17, 5.2°	0.22, 4.0°	0.23, 4.9°	0.30, 12.1°	0.21, 7.8°
	GL-Net [71]*?	8	0.08, 2.8°	0.26, 8.9°	0.17, 11.4°	0.18, 5.1°	0.15, 2.8°	0.25, 4.5°	0.23, 8.8°	0.19, 6.3°
Image based APR	PoseNet [35]*	1	0.32, 6.6°	0.47, 14.0°	0.30, 12.2°	0.48, 7.2°	0.49, 8.1°	0.58, 8.3°	0.48, 13.1°	0.45, 9.9°
	Bayesian PoseNet [33]*	1	0.37, 7.2°	0.43, 13.7°	0.31, 12.0°	0.48, 8.0°	0.61, 7.1°	0.58, 7.5°	0.48, 13.1°	0.47, 9.8°
	Geometric PoseNet [34]*	1	0.13, 4.5°	0.27, 11.3°	0.17, 13.0°	0.19, 5.6°	0.26, 4.8°	0.23, 5.4°	0.35, 12.4°	0.23, 8.1°
	MLFBPPose [67]*	1	0.12, 5.8°	0.26, 12.0°	0.14, 13.5°	0.18, 8.2°	0.21, 7.1°	0.22, 8.1°	0.26, 13.6°	0.20, 9.8°
	Hourglass [45]*	1	0.15, 6.2°	0.27, 10.8°	0.19, 11.6°	0.21, 8.5°	0.25, 7.0°	0.27, 10.2°	0.29, 12.5°	0.23, 9.5°
	LSTM-Pose [64]*	1	0.24, 5.8°	0.34, 11.9°	0.21, 13.7°	0.30, 8.1°	0.33, 7.0°	0.37, 8.8°	0.40, 13.7°	0.31, 9.9°
	BranchNet [68]*	1	0.18, 5.2°	0.34, 9.0°	0.20, 14.2°	0.30, 7.1°	0.27, 5.1°	0.33, 7.4°	0.38, 10.3°	0.29, 8.3°
	ANNet [12]*	1	0.12, 4.3°	0.27, 11.6°	0.16, 12.4°	0.19, 6.8°	0.21, 5.2°	0.25, 6.0°	0.28, 8.4°	0.21, 7.9°
	GPoseNet [13]*	1	0.20, 7.1°	0.38, 12.3°	0.21, 13.8°	0.28, 8.8°	0.37, 6.9°	0.35, 8.2°	0.37, 12.5°	0.31, 10.0°
	AttLoc [65]*	1	0.10, 4.1°	0.25, 11.4°	0.16, 11.8°	0.17, 5.3°	0.21, 4.4°	0.23, 5.4°	0.26, 10.5°	0.20, 7.6°
	AnchorPoint [49]*?	1	**0.06**, 3.9°	**0.16**, 11.1°	**0.09**, 11.2°	**0.11**, 5.4°	**0.14**, 3.6°	**0.13**, 5.3°	**0.21**, 11.9°	0.13, 7.5°
IR	DenseVLAD [59]	1	0.21, 12.5°	0.33, 13.8°	0.15, 14.9°	0.28, 11.2°	0.31, 11.2°	0.30, 11.3°	0.25, 12.3°	0.26, 12.5°
	DenseVLAD+Inter [55]	1	0.18, 10.0°	0.33, 12.4°	0.14, 14.3°	0.25, 10.1°	0.26, 9.4°	0.27, 11.1°	0.24, 14.7°	0.24, 11.7°
RPR	NN-Net [38]	1	0.13, 6.5°	0.26, 12.7°	0.14, 12.3°	0.21, 7.4°	0.24, 6.4°	0.24, 8.0°	0.27, 11.8°	0.21, 9.3°
	RelocNet [3]	1	0.12, 4.1°	0.26, 10.4°	0.14, 10.5°	0.18, 5.3°	0.26, 4.2°	0.23, 5.1°	0.28, 7.5°	0.21, 6.7°
	EssNet [73]	1	0.13, 5.1°	0.27, 10.1°	0.15, 9.9°	0.21, 6.9°	0.22, 6.1°	0.23, 6.9°	0.32, 11.2°	0.22, 8.0°
	EssNet [73] reprod.	1	-	-	-	-	-	-	0.32, 9.8°	-
	NC-EssNet [73]	1	0.12, 5.6°	0.26, 9.6°	0.14, 10.7°	0.20 , 6.7°	0.22, 5.7°	0.22, 6.3°	0.31, 7.9°	0.21, 7.5°
	NC-EssNet [73] reprod.	1	0.13, 5.5°	-	-	-	-	-	-	-
	CamNet [23]?	1	-	-	-	-	-	-	-	**0.05**, **1.8°**
	RelPose GNN	1	0.08, **2.7°**	0.21, **7.5°**	0.13, **8.7°**	0.15, **4.1°**	0.15, **3.5°**	0.19, **3.7°**	0.22, **6.5°**	0.16, 5.2°

"Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision", Turkoglu et al., 3DV'21

SIFT+5Pt
0.08, 2.0°

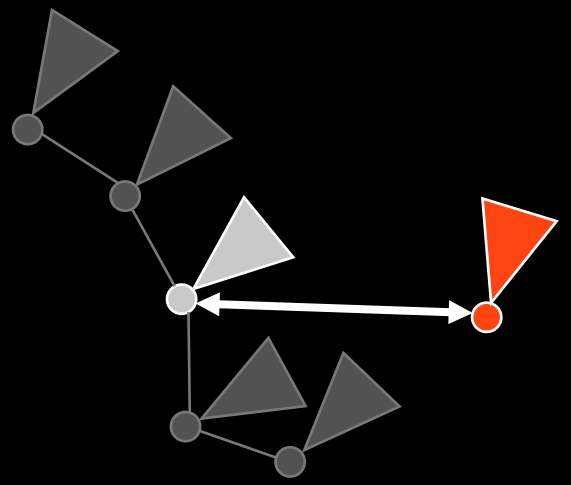
"To Learn or Not to Learn: Visual Localization from Essential Matrices", Zhou at al., ICRA'20

		Matterport-A						Matterport-B					
		R			t			R			t		
		mean (°)	med (°)	rank	mean (°)	med (°)	rank	mean (°)	med (°)	rank	mean (°)	med (°)	rank
DirectionNet	9D	**3.96**	2.28	**2.76**	**14.17**	**6.46**	**3.29**	13.60	**3.54**	**2.89**	**21.26**	**8.90**	**3.44**
	6D	4.30	**2.22**	2.79	16.37	7.07	3.29	14.85	3.69	3.45	23.60	9.42	3.79
	9D-Single	4.55	3.11	3.83	21.65	10.53	4.71	**13.37**	4.00	2.85	28.41	13.27	4.26
	Quat.	23.32	23.00	8.25	39.85	24.85	6.22	37.09	25.25	7.13	49.39	31.59	6.94
Regression	Bin&Delta	6.93	4.71	5.28	22.84	10.16	3.73	31.54	22.98	6.45	29.45	14.30	5.14
	Spherical	10.68	7.98	6.79	40.09	22.85	6.36	32.94	20.56	6.42	51.00	33.18	8.40
	6D	5.73	3.66	3.79	35.75	21.89	6.32	18.23	7.69	4.29	39.06	25.07	5.69
	Quat.	15.40	12.66	6.86	41.57	21.47	7.18	28.38	19.23	6.19	48.99	34.94	7.63
SIFT	LMedS	25.55	5.63	7.71	35.53	14.84	6.20	36.58	10.54	8.13	42.67	26.64	6.06
	RANSAC	19.33	6.66	7.31	45.04	29.78	8.08	31.30	9.55	7.74	47.74	26.19	6.19

Chen et al., “Wide-Baseline Relative Camera Pose Estimation with Directional Learning”, CVPR’21

Relative Pose Regression

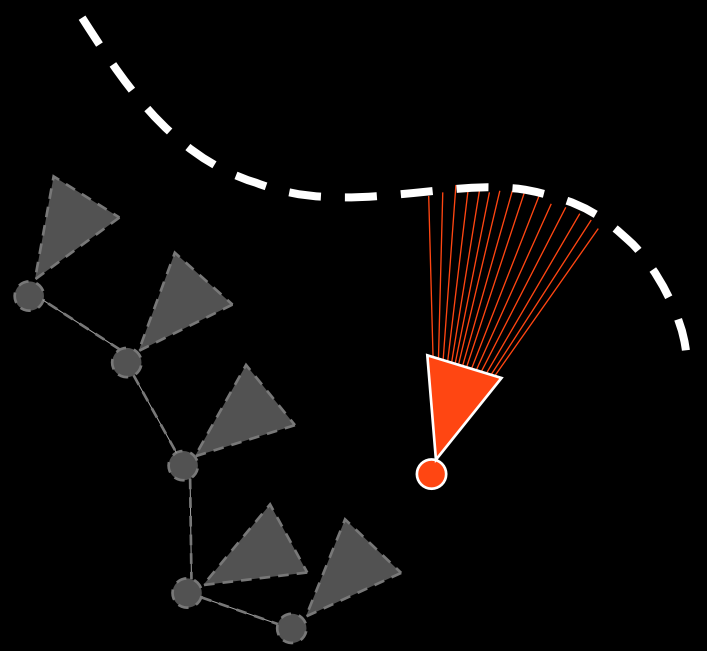
The promise.



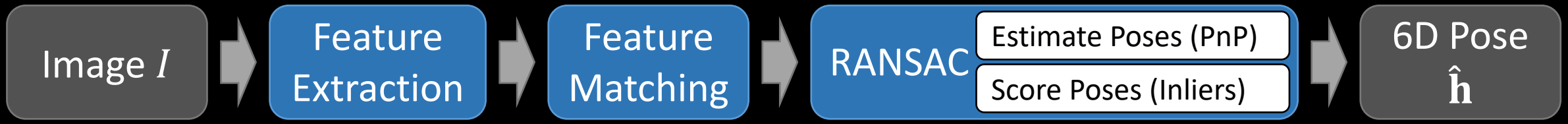
- Works better than APR
- Virtually no mapping costs
- Generalization poor so far

Scene Coordinate Regression

The compromise.



- What is scene coordinate regression?
- Using Random forests
- Using CNNs
- Limits of Benchmarking



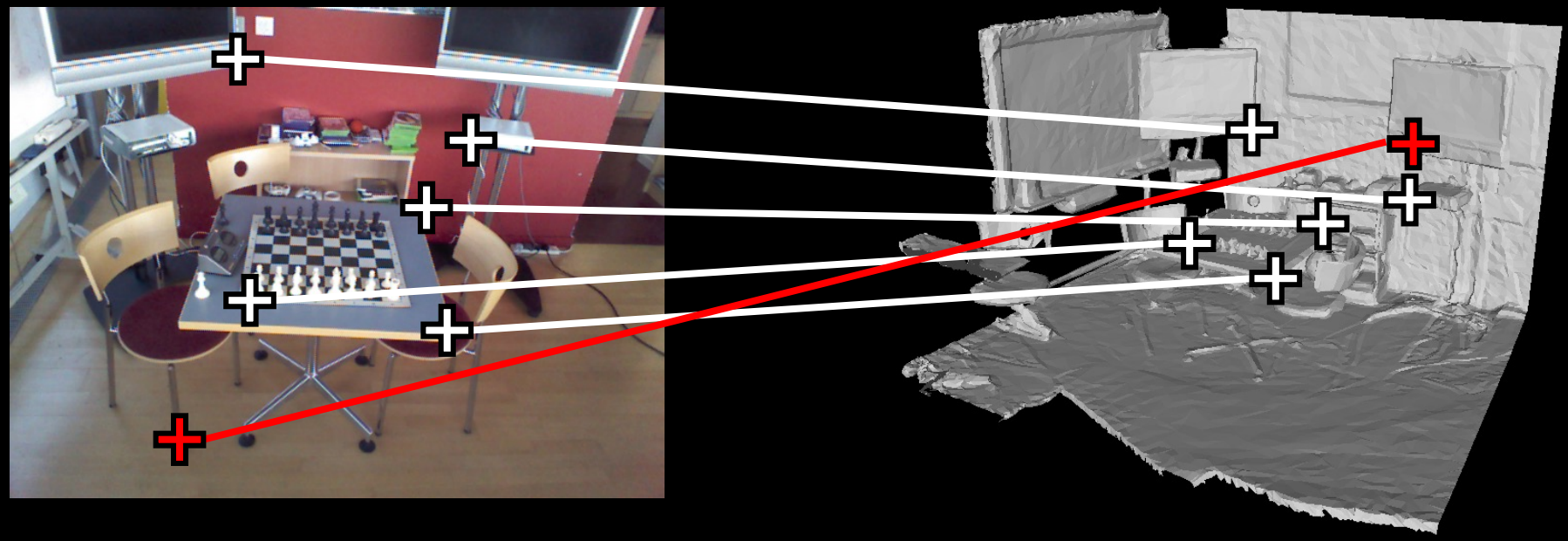
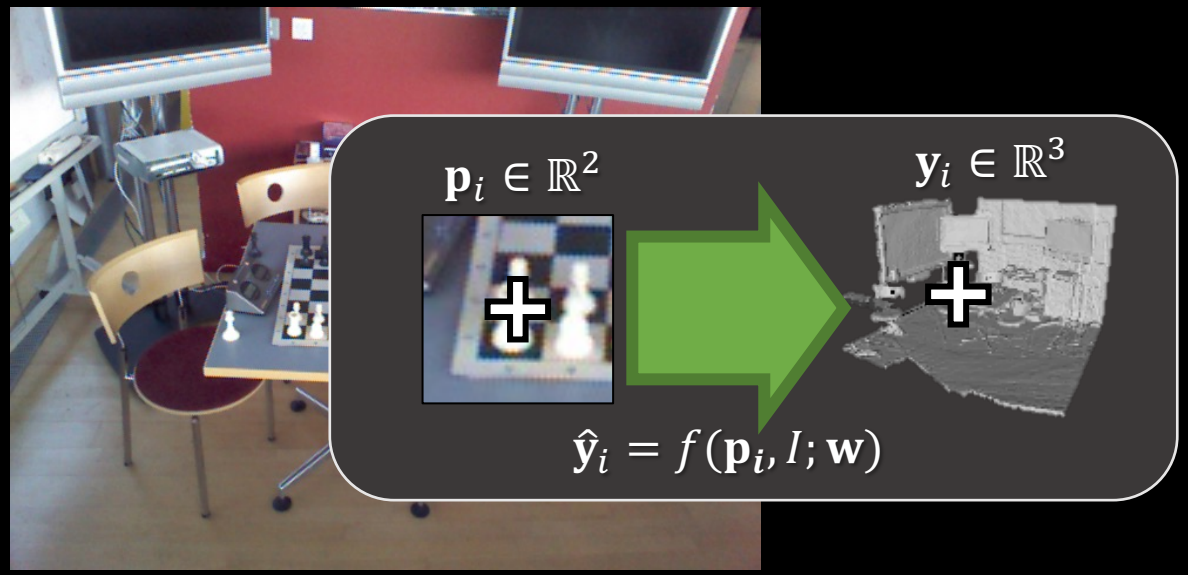
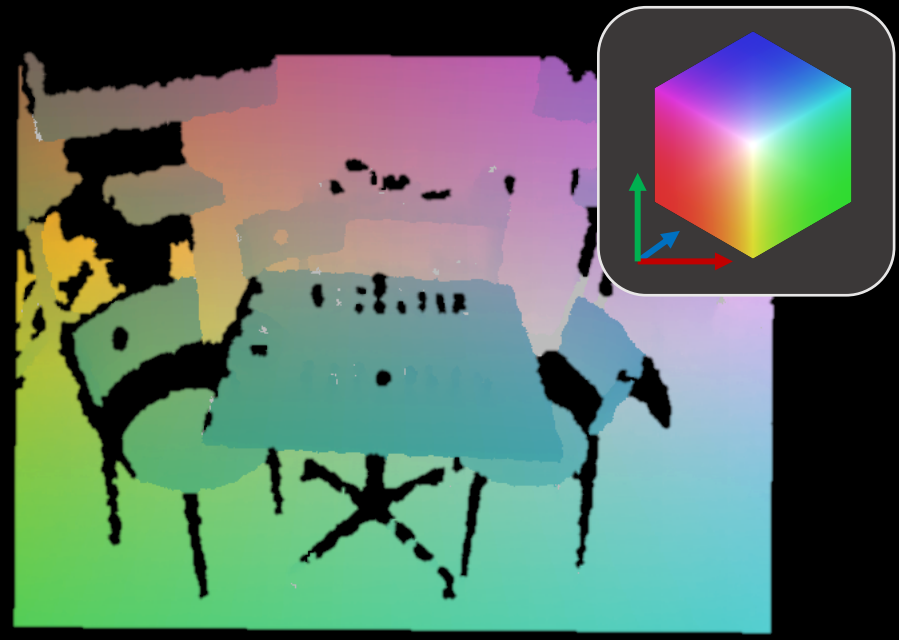


Image Coordinates $\mathbf{p}_i$

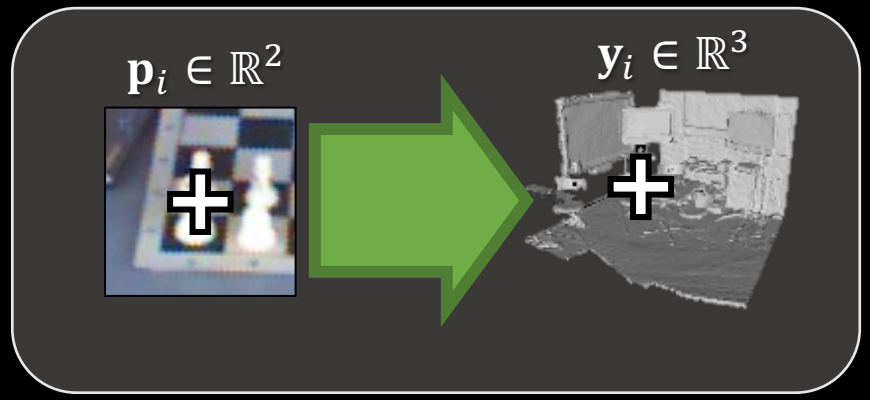
Scene Coordinates $\mathbf{y}_i$

Image I (RGB-D) → Scene Coordinate Regression → RANSAC (Estimate Poses (Kabsch); Score Poses (Inliers)) → 6D Pose $\hat{\mathbf{h}}$



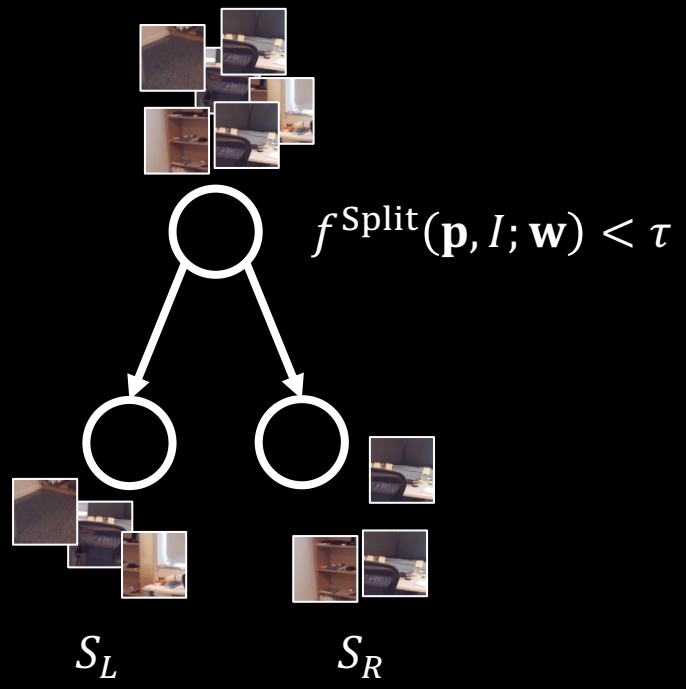


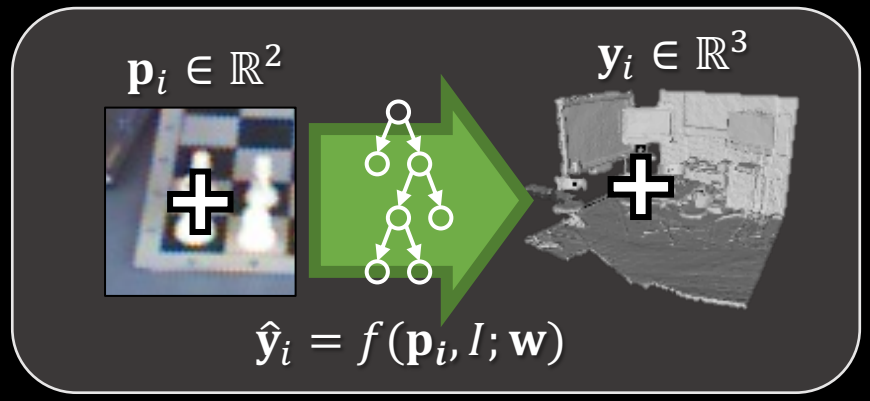
[Sho13] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13



Features: Pixel Differences

Split Score: Reduction of Spatial Variance





Forest Prediction $\hat{\mathbf{y}}_i$

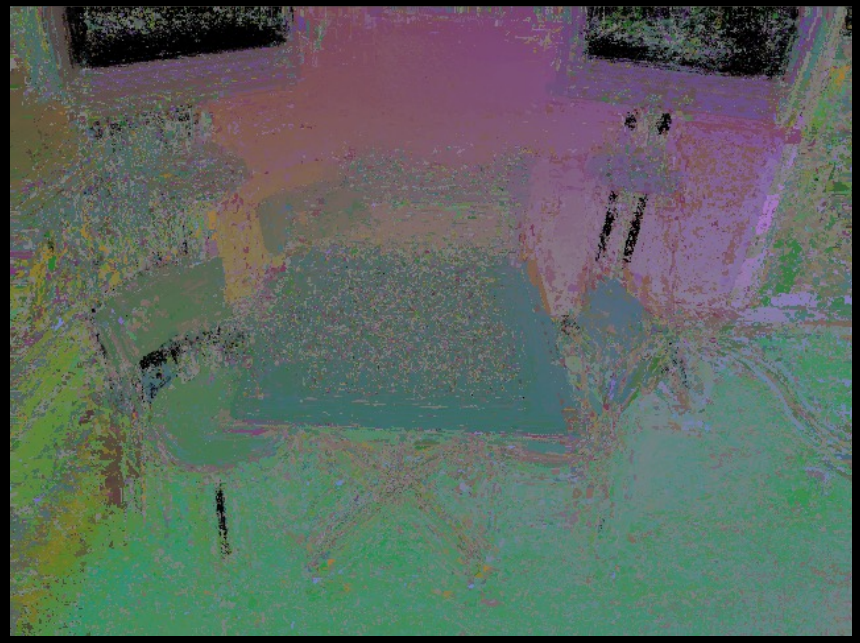
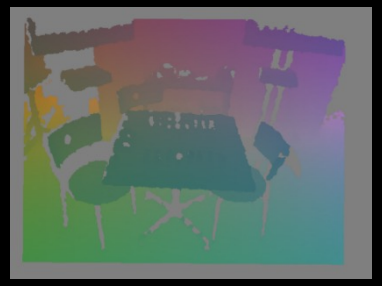
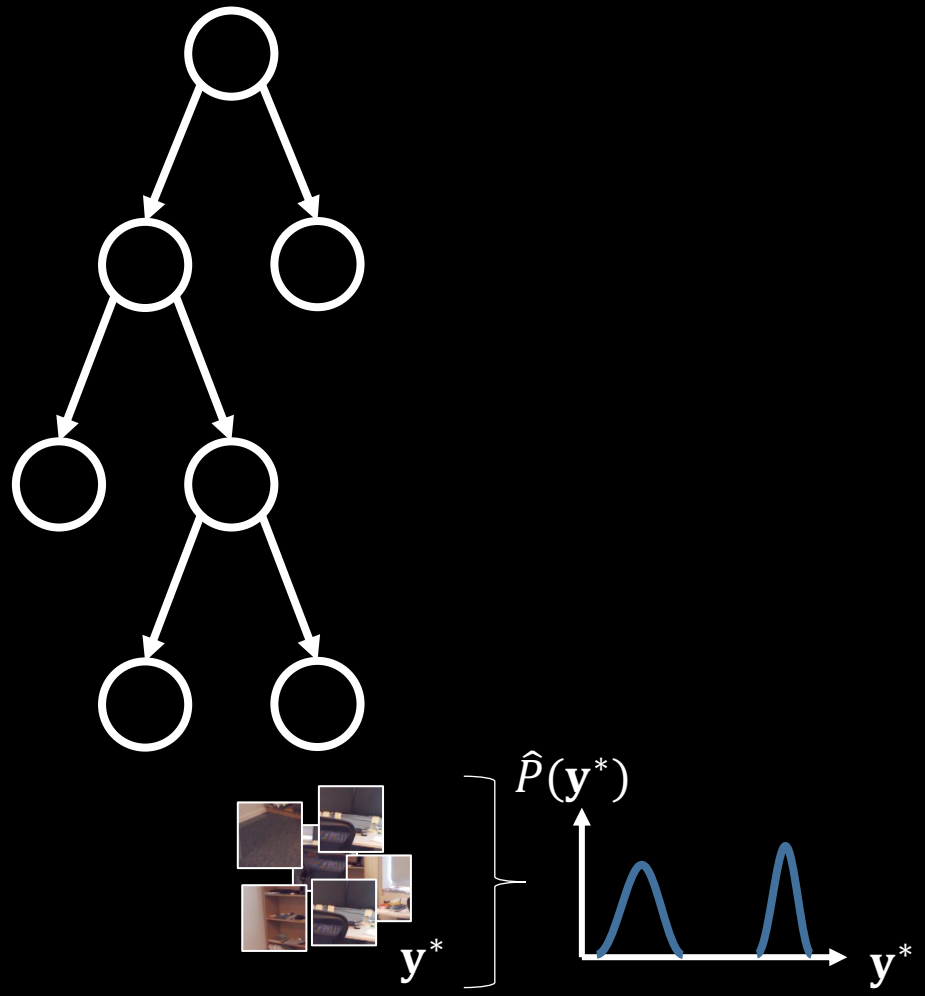
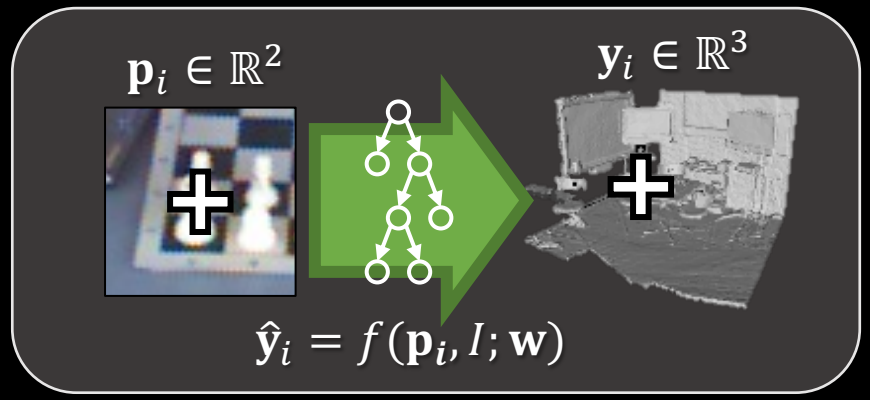


Image I

Ground Truth $\mathbf{y}_i^*$







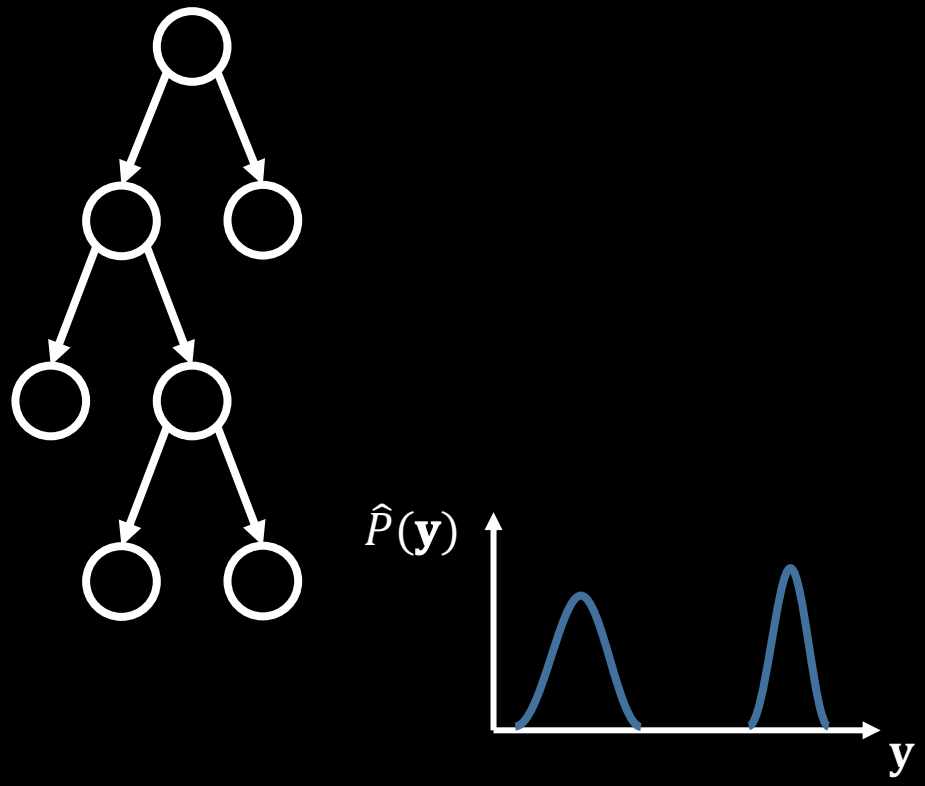
RGB	%Pose Err. $< 5\text{cm}5^\circ$
SCoRF [Bra16]	55.2%
Active Search results from [Bra21]	68.7%
hLoc results from [Bra21]	76.8%
RGB-D	
SCoRF [Sho13]	72.6%
SCoRF [Val15]	**89.2%**

[Sho13] "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images", Shotton et al., CVPR'13
[Val15] "Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization", Valentin et al., CVPR'15
[Bra16] "Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image", Brachmann et al., CVPR'16
[ActiveSearch] "Efficient & effective prioritized matching for large-scale image-based localization", Sattler et al., TPAMI'17
[hLoc] "From Coarse to Fine: Robust Hierarchical Localization at Large Scale", Sarlin et al., CVPR'19
[Bra21] "On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann et al., ICCV'21

Preparation
Scene-Agnostic Training

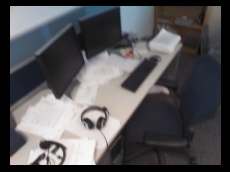
Mapping
Scene-Specific Training

Re-Localisation
Evaluation



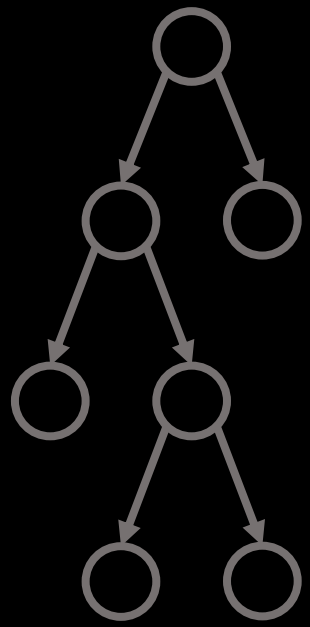
[Cav17] “On-the-Fly Adaptation of Regression Forests for Online Camera Localization”, Cavallari et al., CVPR’17
[Cav19a] “Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade”, Cavallari et al., TPAMI’19
[Cav19b] “Let’s take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation”, Cavallari et al., 3DV’19

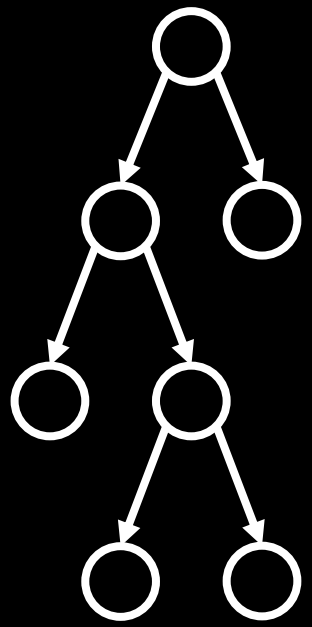
Preparation
Scene-Agnostic Training



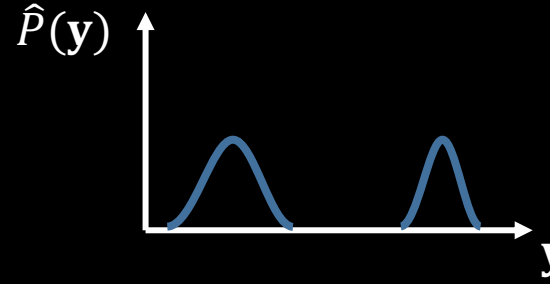
Mapping
Scene-Specific Training

Re-Localisation
Evaluation



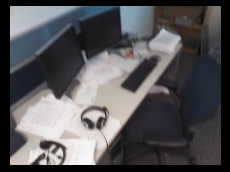


$\hat{\mathbf{y}}$



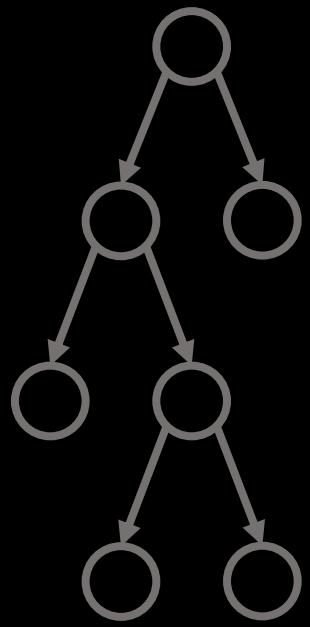
[Cav17] “On-the-Fly Adaptation of Regression Forests for Online Camera Localization”, Cavallari et al., CVPR’17
[Cav19a] “Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade”, Cavallari et al., TPAMI’19
[Cav19b] “Let’s take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation”, Cavallari et al., 3DV’19

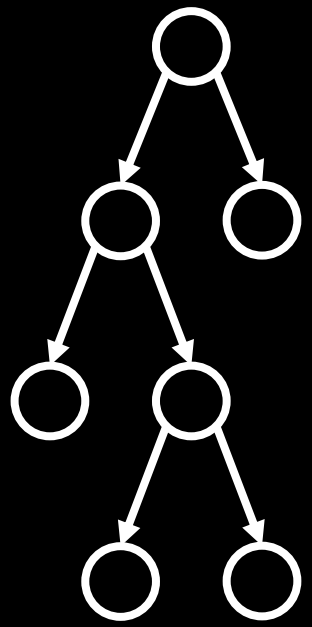
Preparation
Scene-Agnostic Training



Mapping
Scene-Specific Training

Re-Localisation
Evaluation



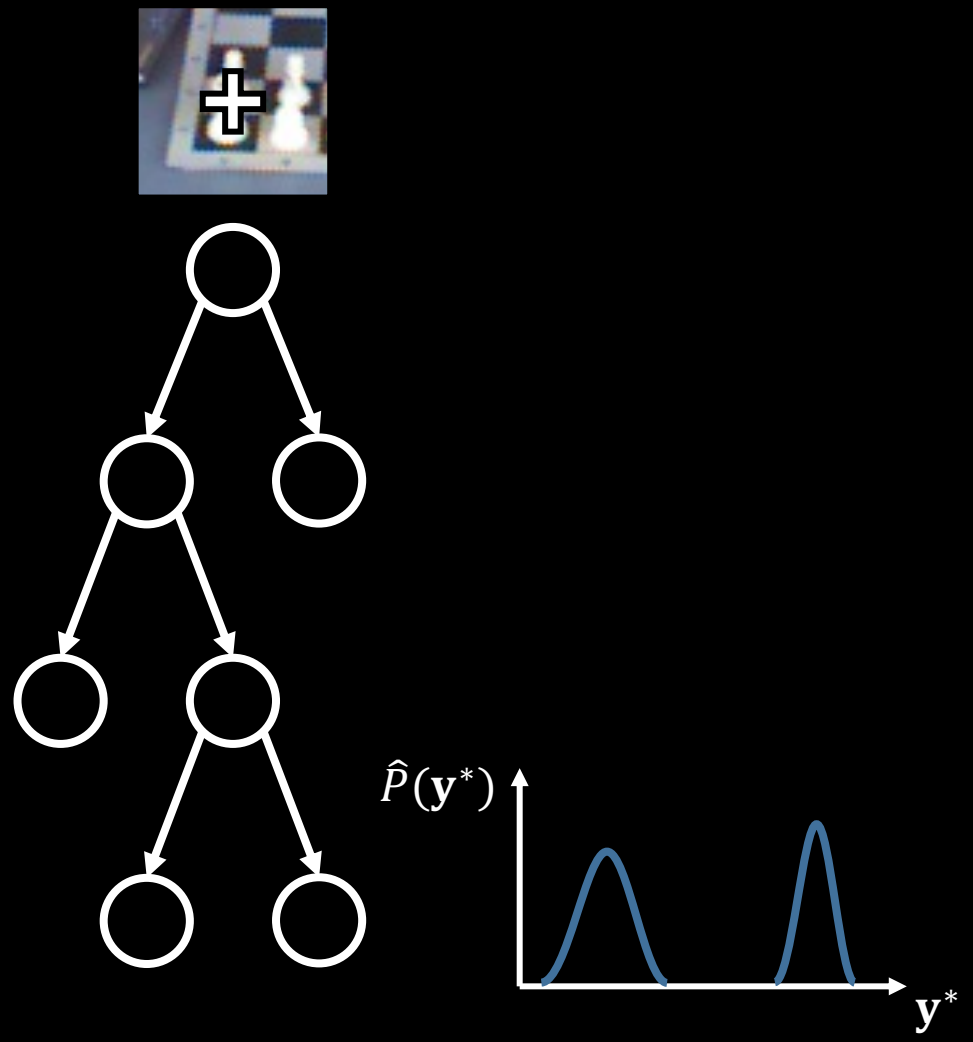


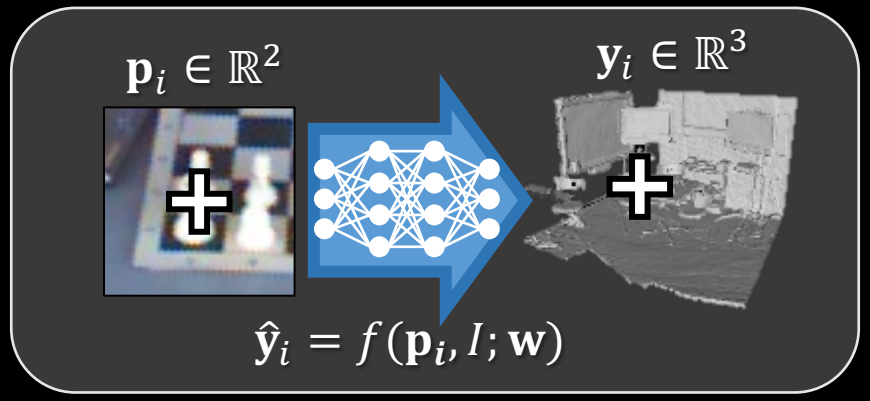
$\hat{P}(\mathbf{y})$

RGB	%Pose Err. $< 5\text{cm}5^\circ$
SCoRF [Bra16]	55.2%
Active Search results from [Bra21]	68.7%
hLoc results from [Bra21]	76.8%
RGB-D	
SCoRF [Sho13]	72.6%
SCoRF [Val15]	89.2%
OtF SCoRF [Cav17]	**96.4%**

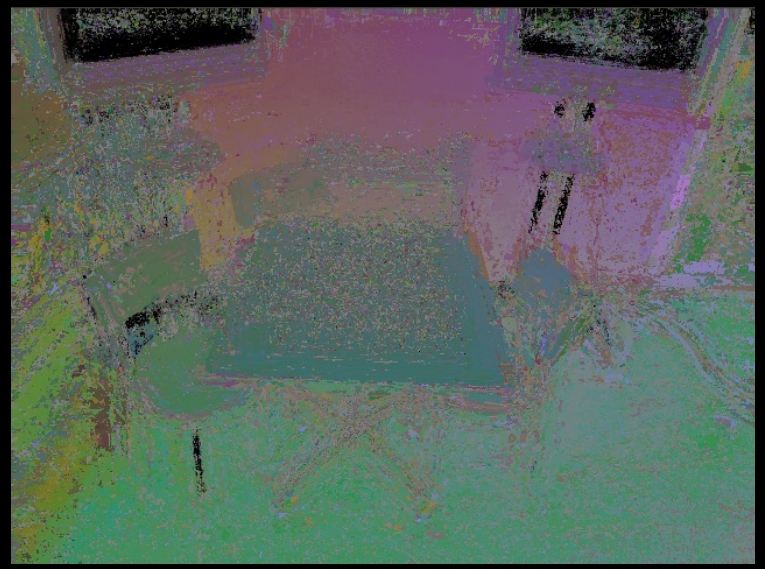
[Cav17] "On-the-Fly Adaptation of Regression Forests for Online Camera Localization", Cavallari et al., CVPR'17
[Cav19a] "Real-time rgb-d camera pose estimation in novel scenes using a relocalisation cascade", Cavallari et al., TPAMI'19
[Cav19b] "Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation", Cavallari et al., 3DV'19

- Advantages:
  - Fast (Training and Inference)
  - Handling Ambiguities
  - Solid Accuracy
- Disadvantages:
  - Needs 3D Model for Training



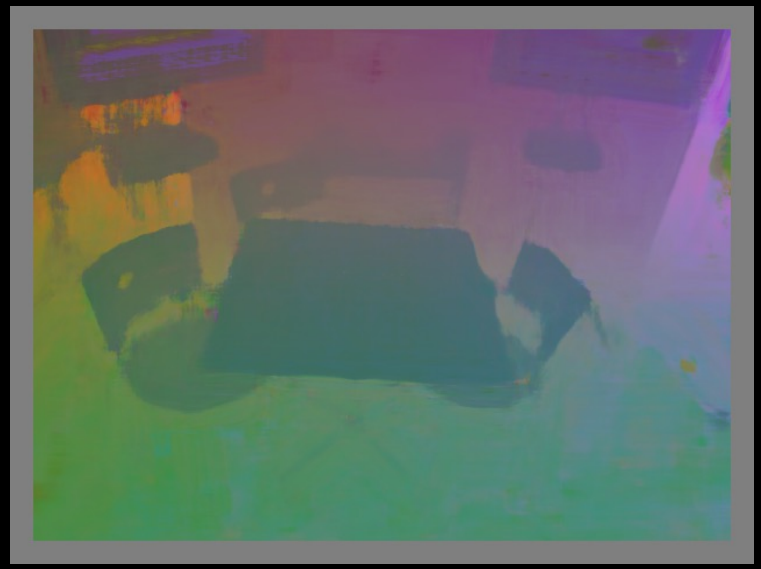


Forest Prediction:



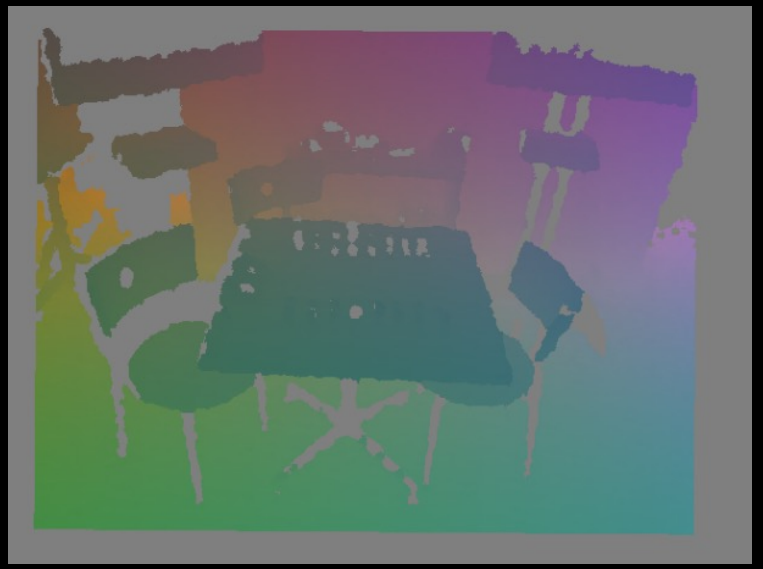
Pose Estimation Succeeds (< 5cm, 5°)

CNN Prediction:



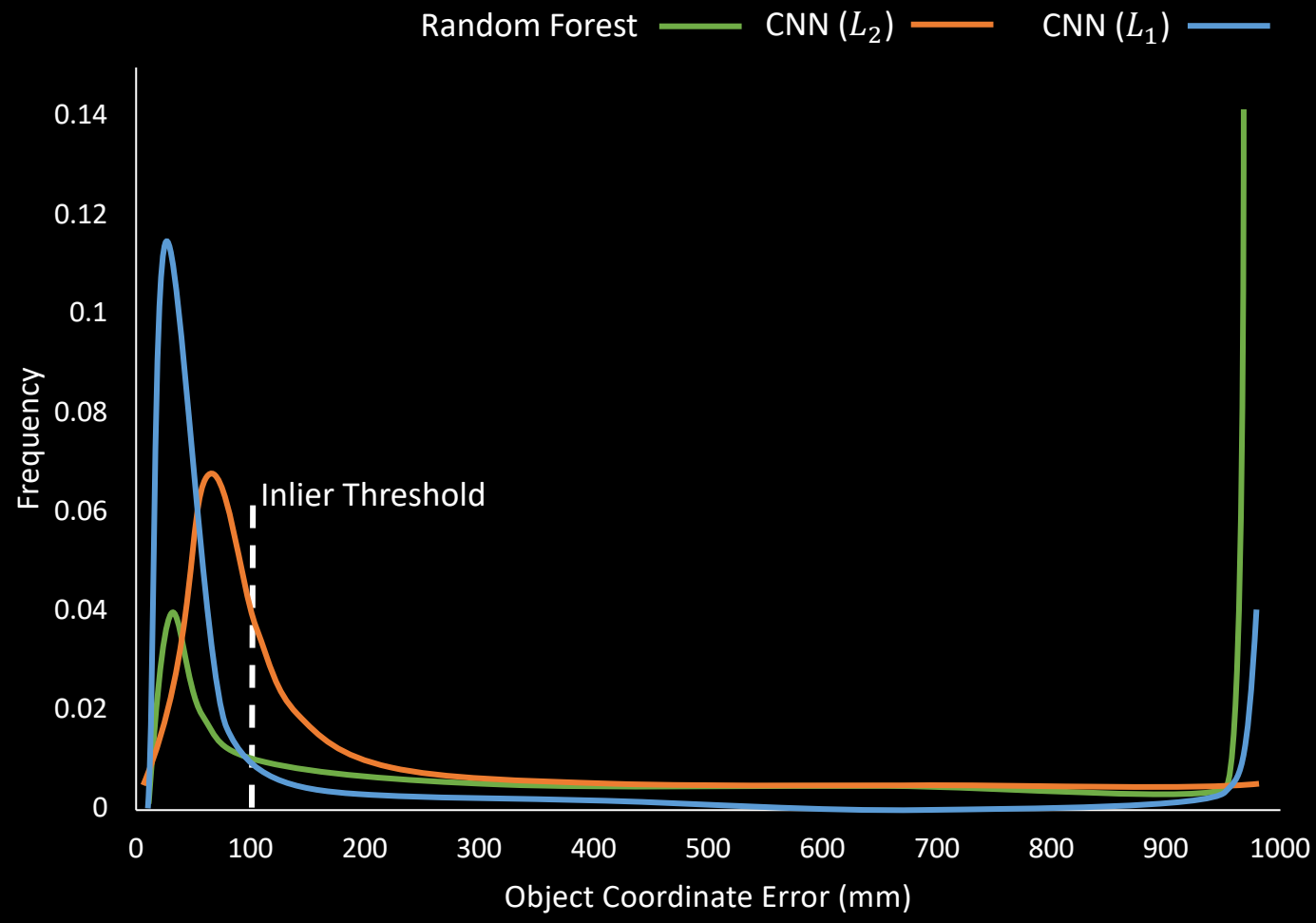
Pose Estimation Fails (> 5cm, 5°)

Ground Truth:



CNN Prediction:

RGB	Scene Coord. Err. $< 10\text{cm}$	%Pose Err. $< 5\text{cm}5^\circ$
Random Forest	~15%	55.2%
CNN (L_2)	~25%	17.2%
CNN (L_1)	~70%	72.8%



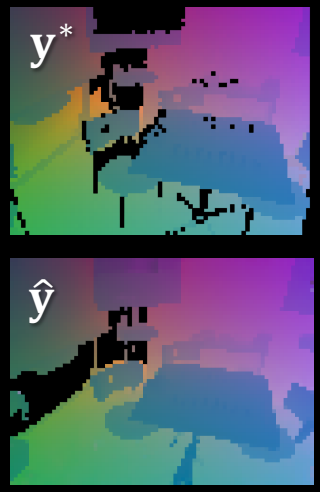
$$\ell_{L_2}(\hat{\mathbf{y}}, \mathbf{y}^*) = \|\hat{\mathbf{y}} - \mathbf{y}^*\|^2$$

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

Reprojection Loss

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

$$\mathcal{V} = \left\{ \mathbf{y} \;\middle|\; \begin{array}{c} \text{in front of camera} \\ \|\hat{\mathbf{y}} - \mathbf{y}^*\| < 0.1\text{m} \\ \text{below maximum reprojection error} \end{array} \right\}$$



$$\ell_{L_1+\pi}(\hat{\mathbf{y}}, \mathbf{h}^*) = \|\mathbf{p} - \pi(\hat{\mathbf{y}}, \mathbf{h}^*)\| \longrightarrow$$

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

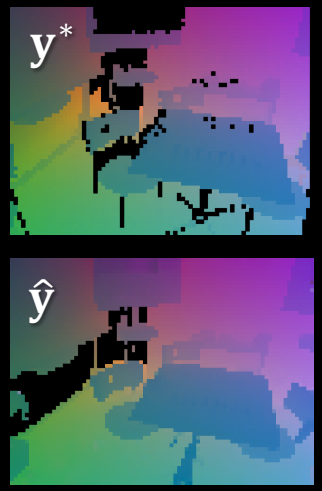
RGB	%Pose Err. $< 5\text{cm}5^\circ$
Active Search results from [Bra21]	68.7%
hLoc results from [Bra21]	76.8%
DSAC*	**80.6%**
RGB-D	
SCoRF [Val15]	89.2%
DSAC*	91.5%
OtF SCoRF [Cav17]	**96.4%**

[Sho13] "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images", Shotton et al., CVPR'13
[Val15] "Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization", Valentin et al., CVPR'15
[Bra16] "Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image", Brachmann et al., CVPR'16
[ActiveSearch] "Efficient & effective prioritized matching for large-scale image-based localization", Sattler et al., TPAMI'17
[hLoc] "From Coarse to Fine: Robust Hierarchical Localization at Large Scale", Sarlin et al., CVPR'19
[Bra21] "On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann et al., ICCV'21
[DSAC*] "Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC", Brachmann and Rother, TPAMI'21

DSAC* Training w/ 3D Model

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

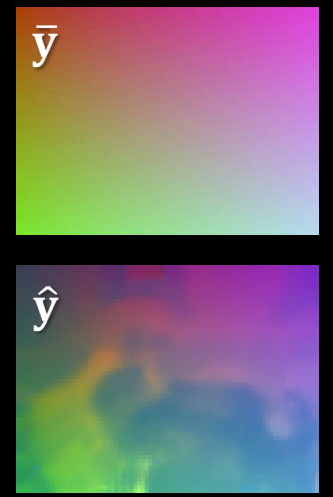
$$\mathcal{V} = \left\{ \mathbf{y} \;\middle|\; \begin{array}{c} \text{in front of camera} \\ \|\hat{\mathbf{y}} - \mathbf{y}^*\| < 0.1\text{m} \\ \text{below maximum reprojection error} \end{array} \right\}$$



DSAC* Training w/o 3D Model

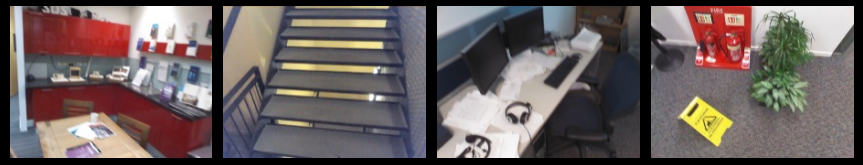
$$\ell_{L_1+\pi}(\hat{\mathbf{y}}, \mathbf{h}^*) = \begin{cases} \|\mathbf{p} - \pi(\hat{\mathbf{y}}, \mathbf{h}^*)\|, & \hat{\mathbf{y}} \in \mathcal{V} \\ \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|, & \text{otherwise} \end{cases}$$

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

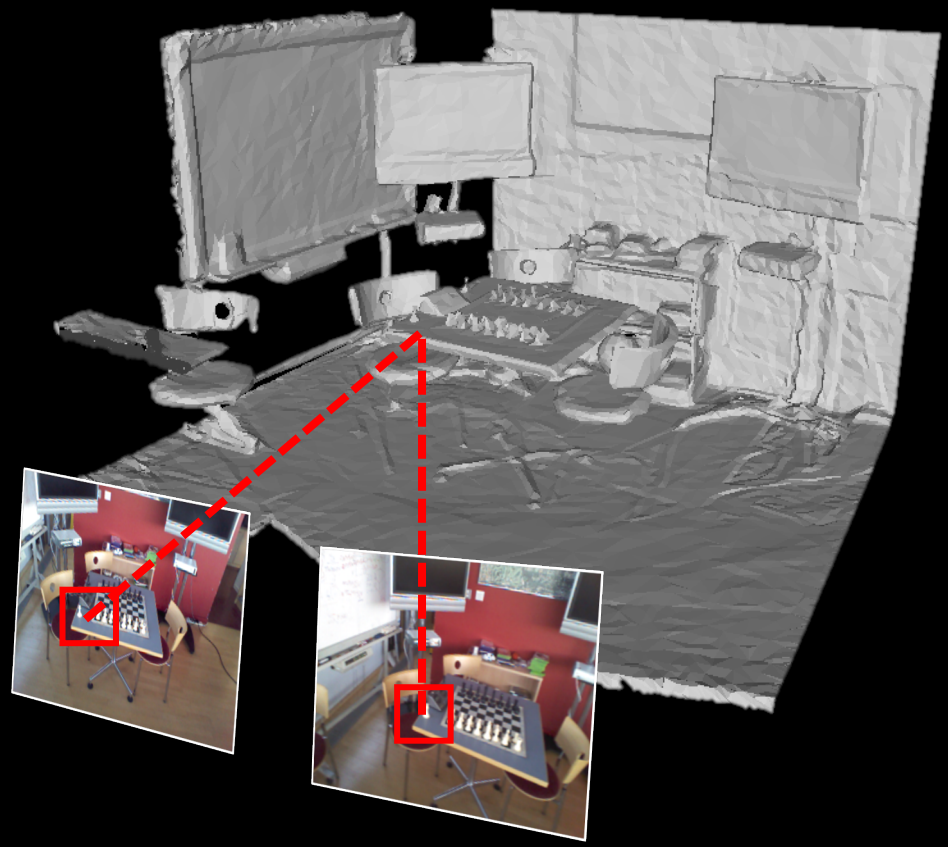


[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21

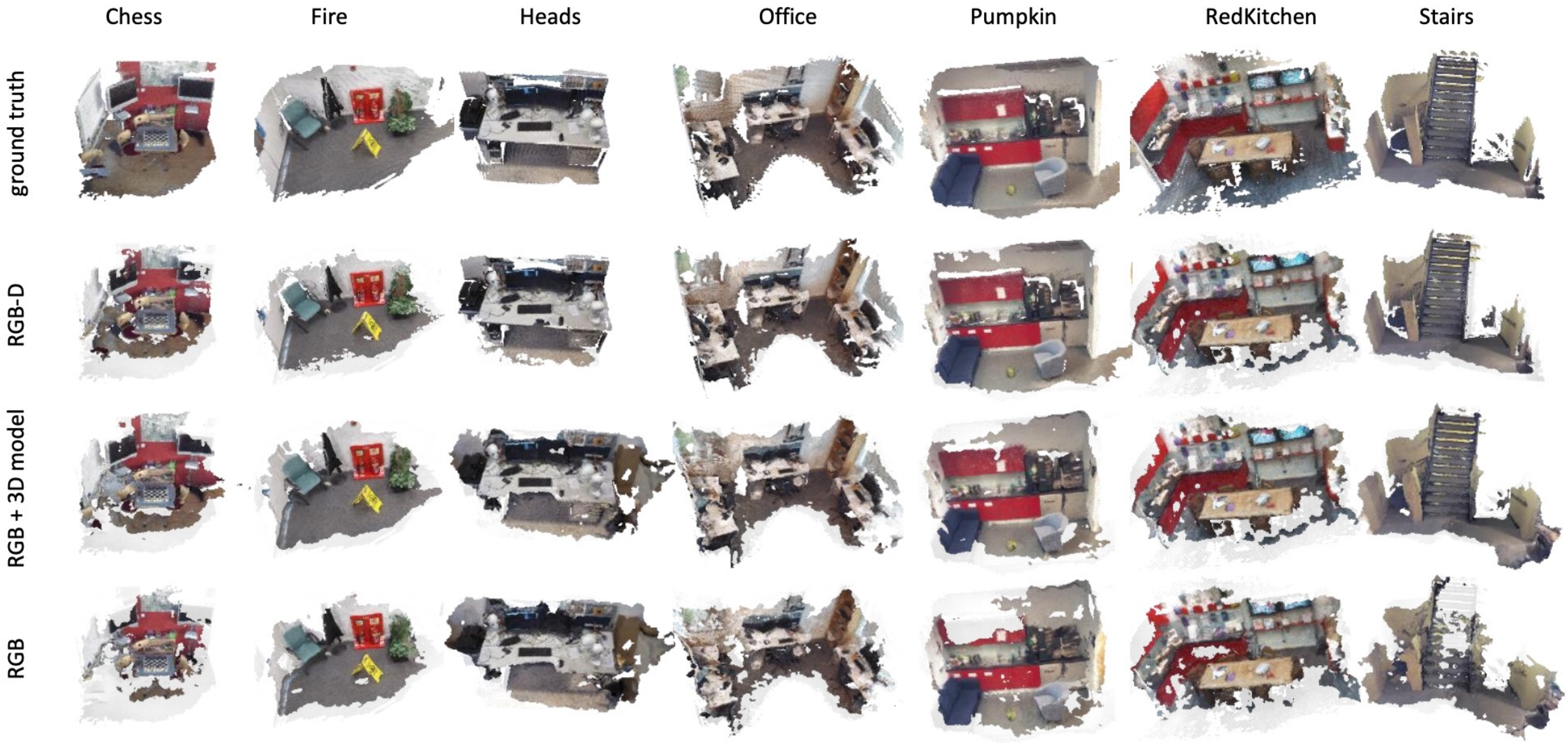
7Scenes Dataset [Sho13]

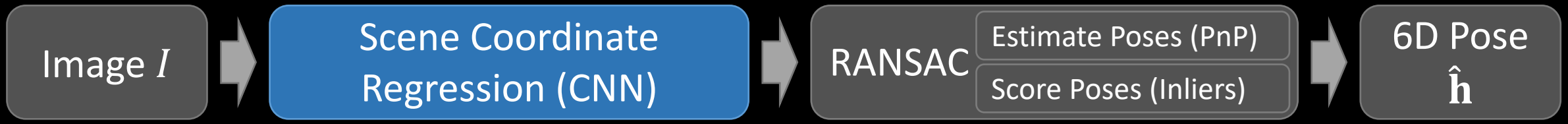


	%Pose Err. $< 5cm5^{\circ}$
[Li18] w/ multi-view	69.4%
hLoc results from [Bra21]	76.8%
[DSAC*] w/o 3D model	80.7%
[DSAC*] w/ 3D model	**85.2%**



[DSAC*] "Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC", Brachmann and Rother, TPAMI'21
[Li18] "Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization", Li et al., ECCV'18 Workshops
[hLoc] "From Coarse to Fine: Robust Hierarchical Localization at Large Scale", Sarlin et al., CVPR'19
[Bra21] "On the Limits of Pseudo Ground Truth in Visual Camera Re-localisation", Brachmann et al., ICCV'21

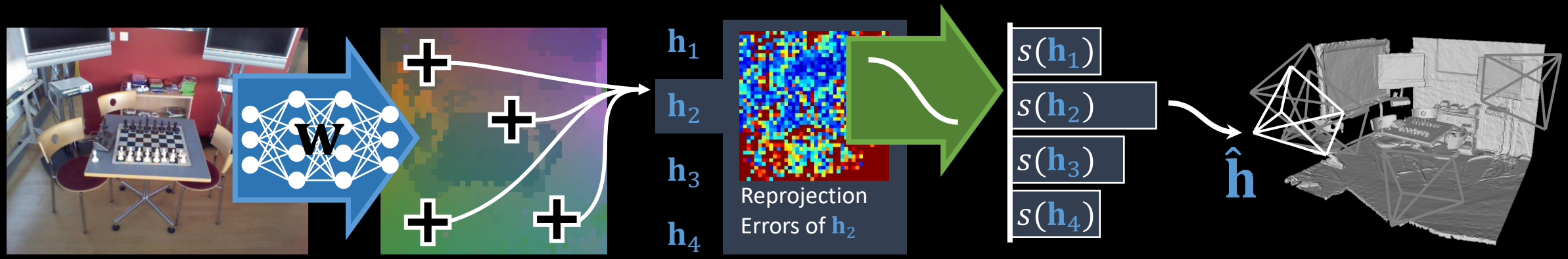

Chess
Fire
Heads
Office
Pumpkin
RedKitchen
Stairs
ground truth
RGB-D
RGB + 3D model
RGB



$$\ell(\hat{\mathbf{y}}, \mathbf{y}^*)$$

$$\ell(\hat{\mathbf{h}}, \mathbf{h}^*)$$

Image I → Scene Coordinate Regression (CNN) → RANSAC (Estimate Poses (PnP); Score Poses (Inliers)) → 6D Pose $\hat{\mathbf{h}}$



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

Differentiating PnP [DSAC++]:

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

Soft Inlier Counting [DSAC++] :

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

argmax Selection

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

[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

argmax Selection

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

Probabilistic Selection [DSAC]

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

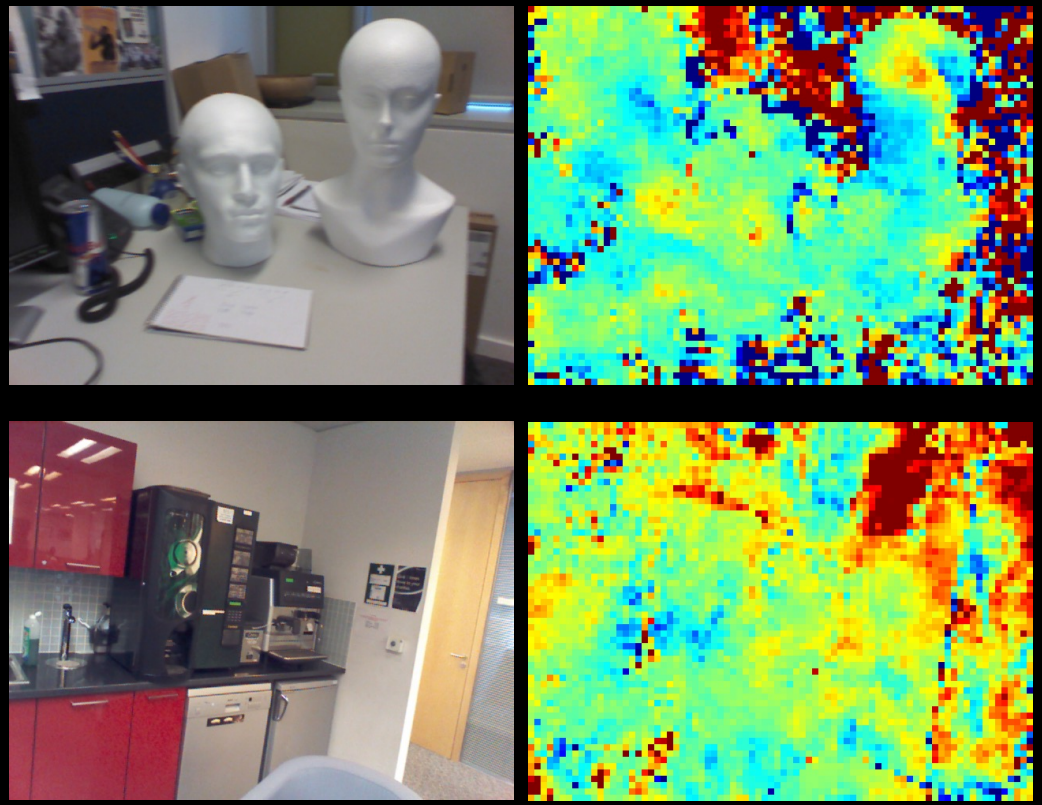
DSAC Learning objective:

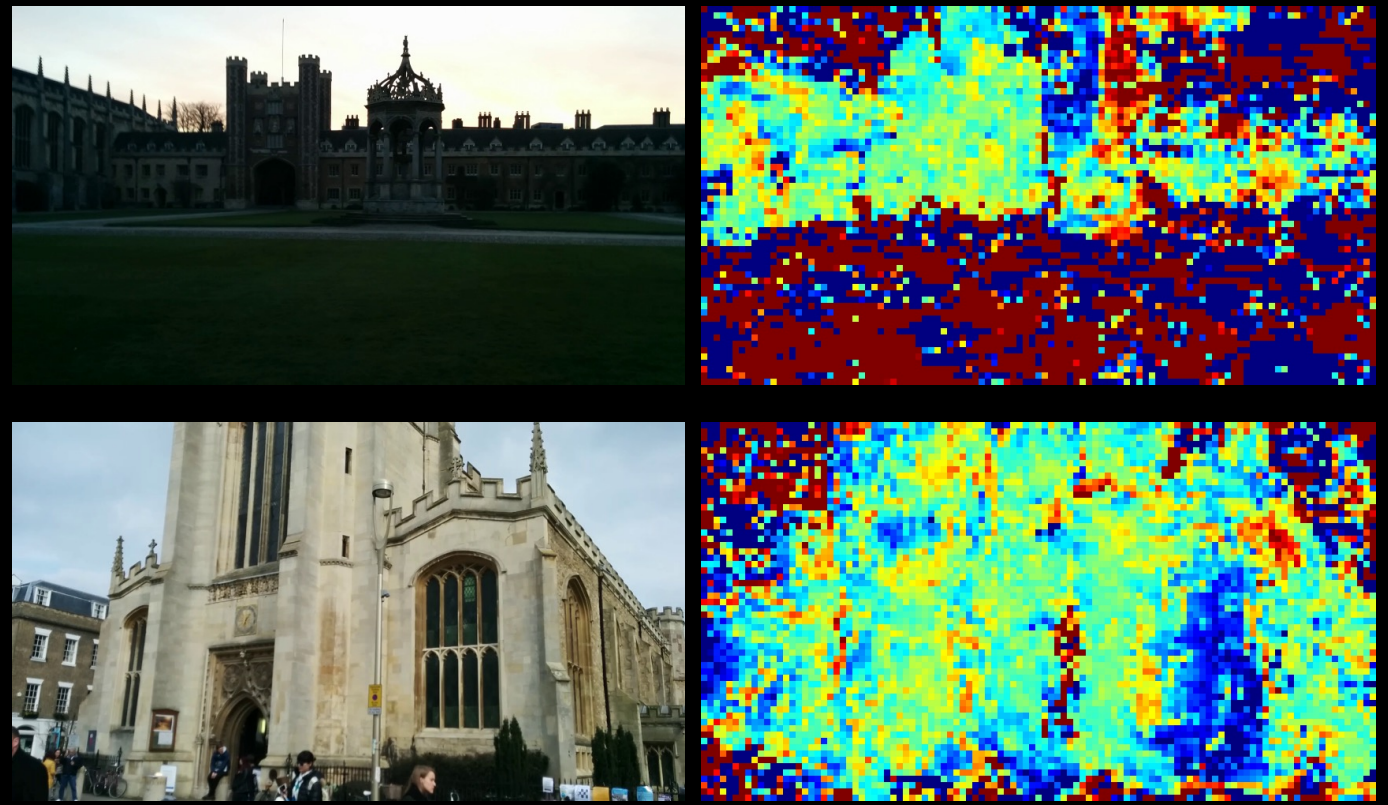
$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{j \sim P(j;\mathbf{w})}\left[\ell(\mathbf{h}_j, \mathbf{h}^*)\right]$$

RGB	**%Pose Err. $< 5\text{cm}5^\circ$**
Active Search results from [Bra21]	68.7%
hLoc results from [Bra21]	76.8%
DSAC* ($\ell_{L_1+\pi}$)	80.6%
DSAC* (ℓ_{DSAC})	**85.2%**
RGB-D	
SCoRF [Val15]	89.2%
DSAC*(ℓ_{L_1})	91.5%
DSAC* (ℓ_{DSAC})	94.6%
OtF SCoRF [Cav17]	**96.4%**

[DSAC] “DSAC - Differentiable RANSAC for camera localization”, Brachmann et al., CVPR’17
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21

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

	Architecture	GPU	Approximate Time Per Frame
Active Search		x	50ms
PoseNet	GoogLeNet	x	5ms
SCoRF	Random Forest		100ms
DSAC++	VGGNet	x	200ms
DSAC*	ResNet	x	75ms (30ms) On Tesla K80 (GeForce GTX 2080 Ti)

[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17
[SCoRF] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13
[DSAC++] “Learning Less is More – 6D Camera Localization via 3D Surface Regression”, Brachmann and Rother, CVPR’18
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21

	Architecture	GPU	Approximate Training Time
Active Search			- (but SfM reconstruction)
PoseNet	GoogLeNet	x	1h
SCoRF	Random Forest		10min
DSAC++	VGGNet	x	$6\mathrm{d} = 2\mathrm{d}\times\ell_{L_1} + 2\mathrm{d}\times\ell_{\pi} + 2\mathrm{d}\times\ell_{\mathbf{h}}$
DSAC*	ResNet	x	$2.5\mathrm{d} = 2\mathrm{d}\times\ell_{L_1+\pi} + 0.5\mathrm{d}\times\ell_{\mathbf{h}}$ **(16h = 8h+8h)** On Tesla K80 (GeForce GTX 2080 Ti)

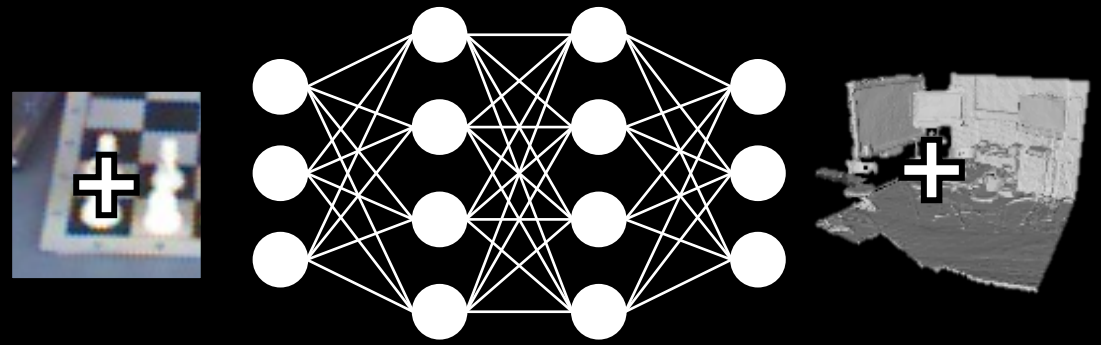
[ActiveSearch] “Efficient & effective prioritized matching for large-scale image-based localization”, Sattler et al., TPAMI’17
[PoseNet] “Geometric Loss Functions for Camera Pose Regression with Deep Learning” Kendall and Cipolla, CVPR ’17
[SCoRF] “Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images”, Shotton et al., CVPR’13
[DSAC++] “Learning Less is More – 6D Camera Localization via 3D Surface Regression”, Brachmann and Rother, CVPR’18
[DSAC*] “Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC”, Brachmann and Rother, TPAMI’21

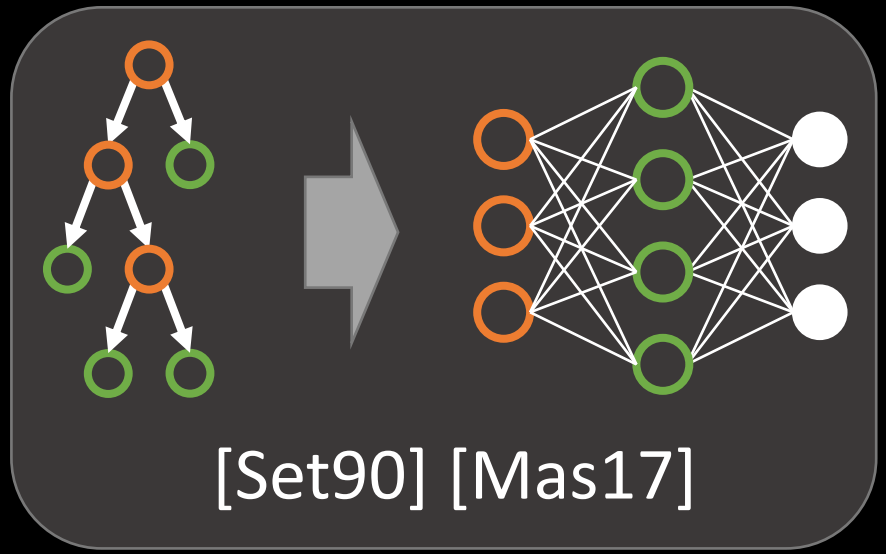
Cambridge Landmarks [Ken15]


	Kings College		Old Hospital		Shop Facade		St Mary's Church	
	Memory Demand	Median Error	Memory Demand	Median Error	Memory Demand	Median Error	Memory Demand	Median Error
ActiveSearch v1.1	177MB	0.14m	94MB	0.20m	38MB	0.05m	226MB	0.09m
HybridCompression	1.0MB	0.81m	0.62MB	0.75m	0.2MB	0.19m	1.3MB	0.50m
DenseVLAD	10.1MB	2.8m	14.0MB	4.0m	3.6MB	1.1m	23.2MB	2.3m
PoseNet	50MB	0.88m	50MB	3.2m	50MB	0.88m	50MB	1.57m
DSAC++ (LUA/Torch, VGGNet)	207MB	0.18m	207MB	0.20m	207MB	0.06m	207MB	0.13m
DSAC++ (PyTorch, VGGNet)	104MB	0.18m	104MB	0.20m	104MB	0.06m	104MB	0.13m
DSAC* (PyTorch, ResNet)	28MB	0.15m	28MB	0.21m	28MB	0.05m	28MB	0.13m
DSAC*-Tiny (PyTorch, ResNet)	3.8MB	0.19m	3.8MB	0.23m	3.8MB	0.07m	3.8MB	0.39m

Best Second Third Fourth Fifth and worse

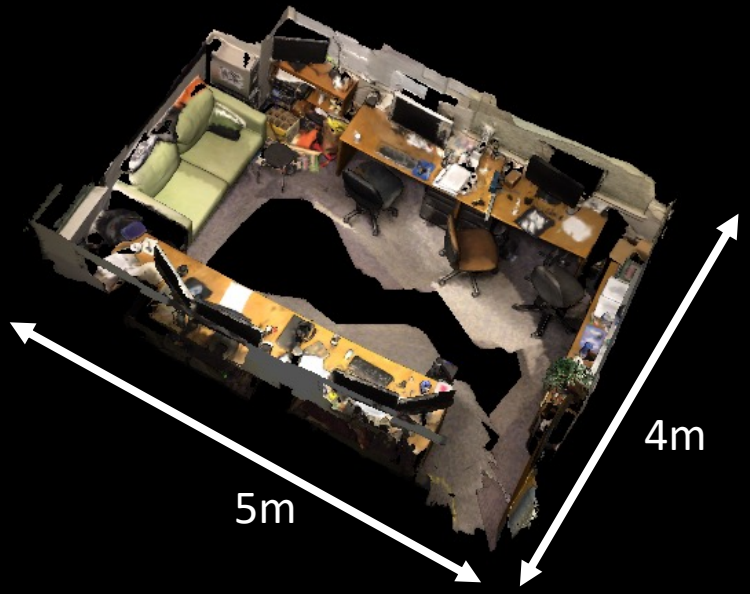
- Advantages:
  - High Accuracy
  - End-To-End Training
  - Training w/o depth or 3D model
  - Good scene compression
- Disadvantages:
  - Slow (esp. training)





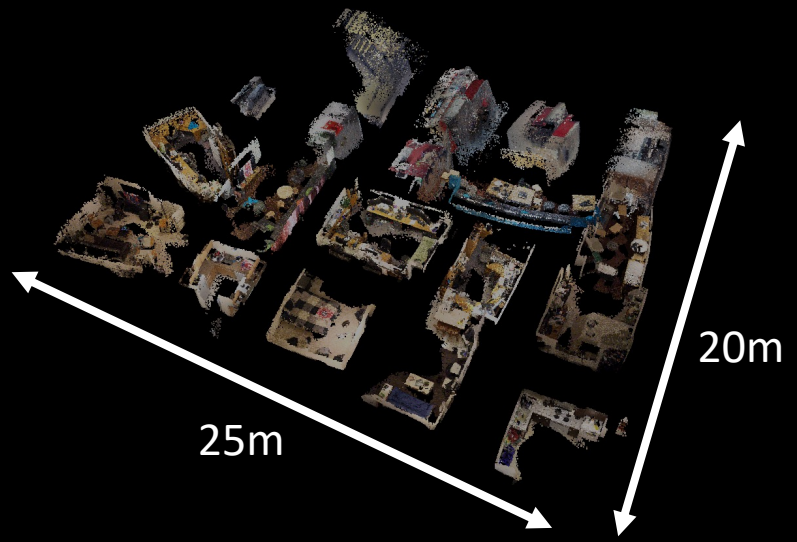
[Set90] „Entropy nets: From decision trees to neural networks”, Sethi, IEEE’90
[Mas17] “Random Forests versus Neural Networks – What’s Best for Camera Localization?”, Massiceti et al., ICRA’17

7Scenes Dataset [Sho13]
12Scenes Dataset [Val16]



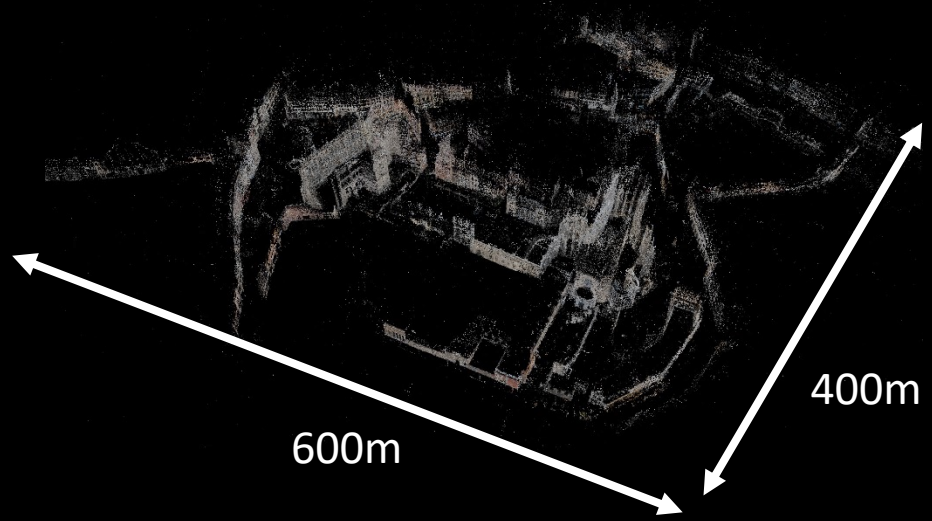
Average Accuracy (5cm,5°):
DSAC++: 88.6%
ORB+PnP: 48.9%

7Scenes+12Scenes [ESAC]



Average Accuracy (5cm,5°):
DSAC++: 53.3%

Aachen Dataset [Sat12]



Average Accuracy (25cm,2°):
DSAC++: 0.4%
ActiveSearch v1.1: 85.3%

[Sho13] "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images", Shotton et al., CVPR'13
[Val16] "Learning to Navigate the Energy Landscape", Valentin et al., 3DV'16
[Sat12] "Image Retrieval for Image-Based Localization Revisited", Sattler et al., BMVC'12
[ESAC] „Expert Sample Consensus Applied to Camera Re-Localization", Brachmann and Rother, ICCV'19

- large-scale re-localization
- training an ensemble of expert networks
- gating network distributes computation budget to experts

Train all networks jointly and end-to-end:

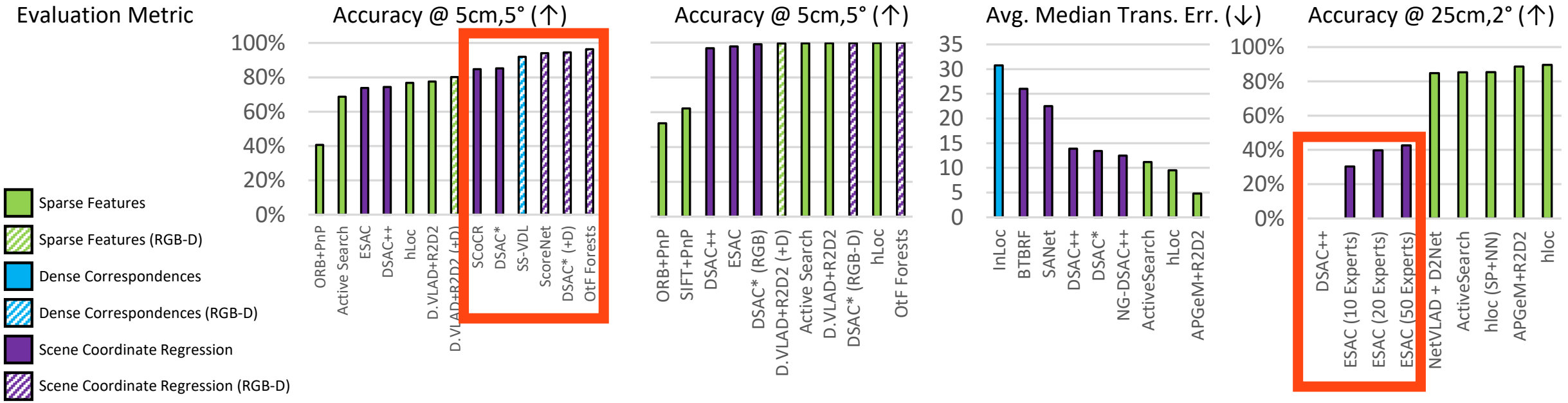
$$\mathcal{L}(\mathbf{w}) = \mathbb{E}_{\mathcal{H}\sim P(\mathcal{H})}\mathbb{E}_{j\sim P(j|\mathcal{H})}\left[\ell(\mathbf{h}_j)\right]$$

Average Accuracy (25cm,2°):
DSAC++: 0.4%
ESAC: 42.6%

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

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

Dataset	7Scenes [1]	12Scenes [2]	Cambridge [3]	Aachen Day [4]



[1] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images". In CVPR, 2013

[2] J. Valentin, A. Dai, M. Niessner, P. Kohli, P. Torr, S. Izadi, and C. Keskin. "Learning to Navigate the Energy Landscape". In 3DV, 2016

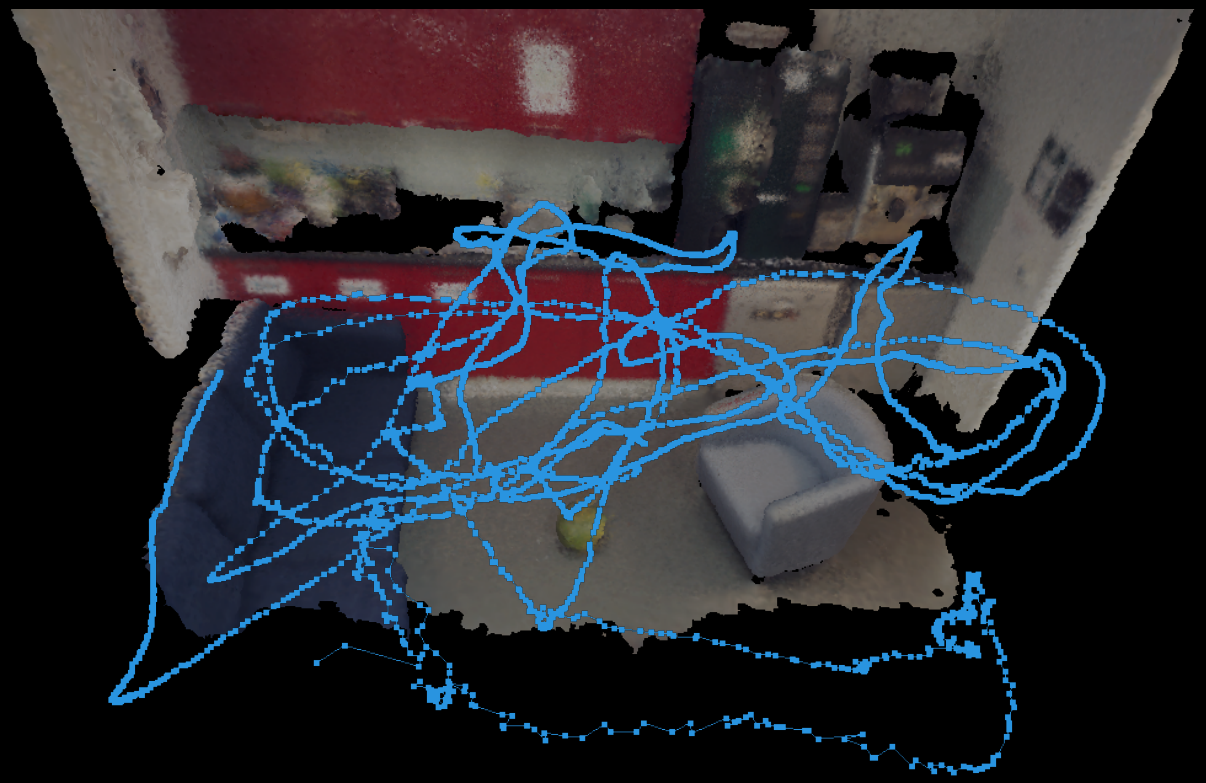
[3] A. Kendall, M. Grimes, and R. Cipolla. "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization". In ICCV, 2015

[4] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. "Benchmarking 6DOF Urban Visual Localization in Changing Conditions". In CVPR, 2018

D-SLAM
(Kinect Fusion)

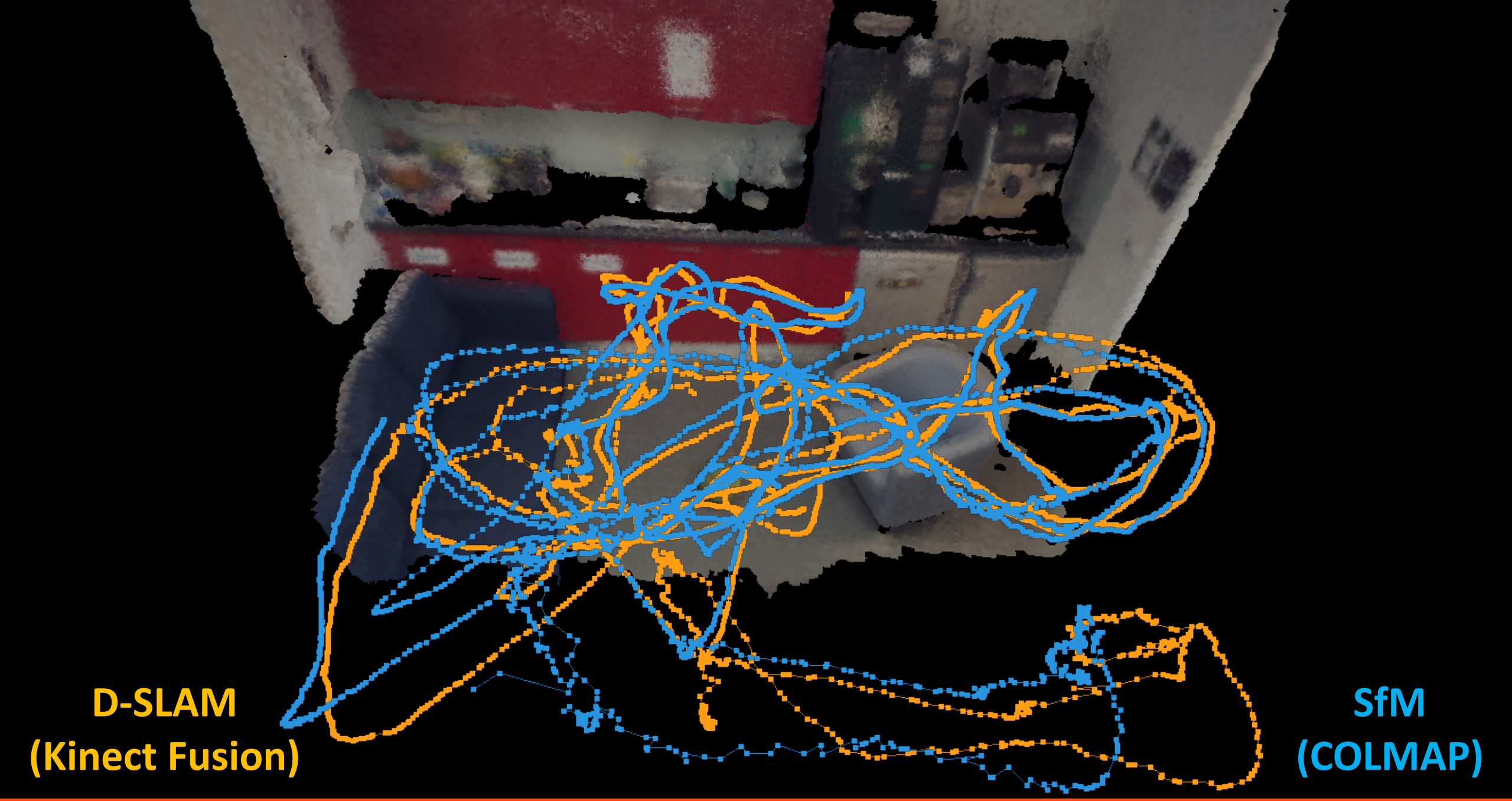
SfM
(COLMAP)

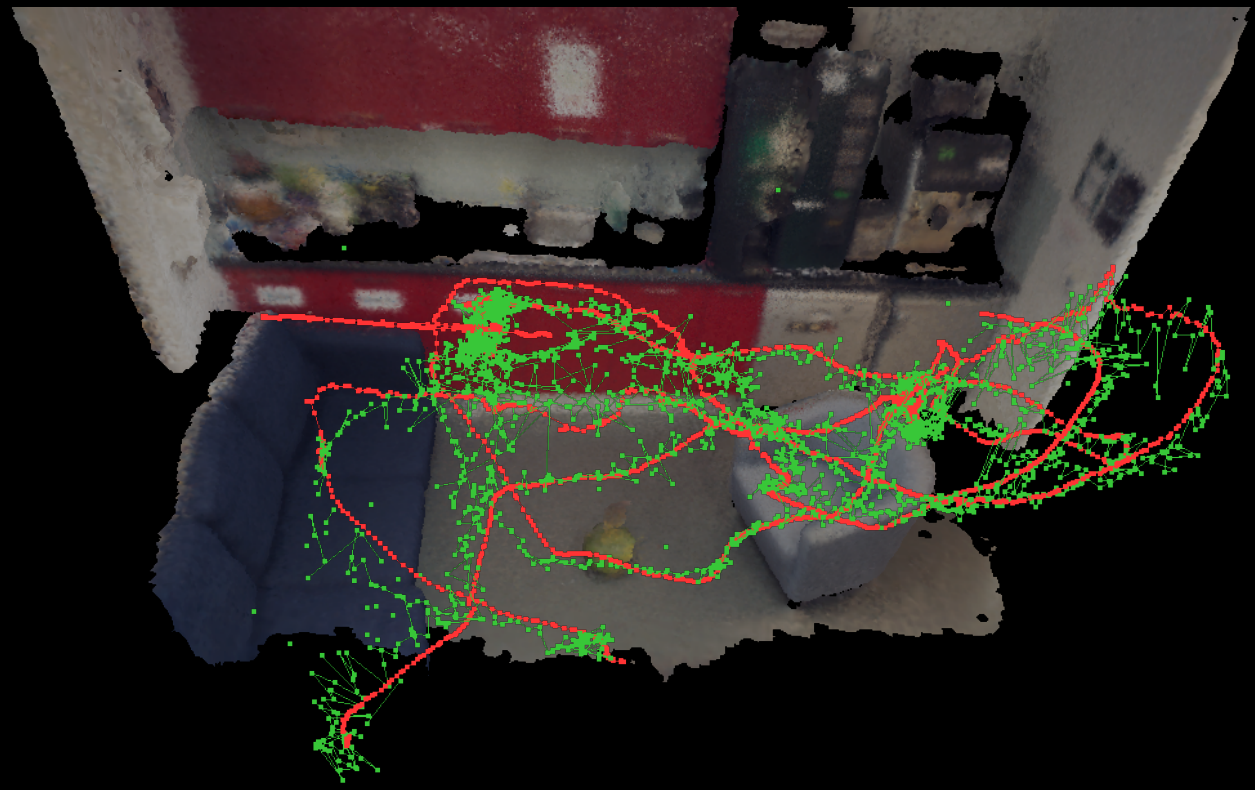
R. A. Newcombe, S. Izadi, O. Hilliges, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. “KinectFusion: Real-Time Dense Surface Mapping and Tracking”. In ISMAR, 2011
J. L. Schönberger and J.-M. Frahm. “Structure-From-Motion Revisited”. In CVPR, 2016



D-SLAM
(Kinect Fusion)

SfM
(COLMAP)

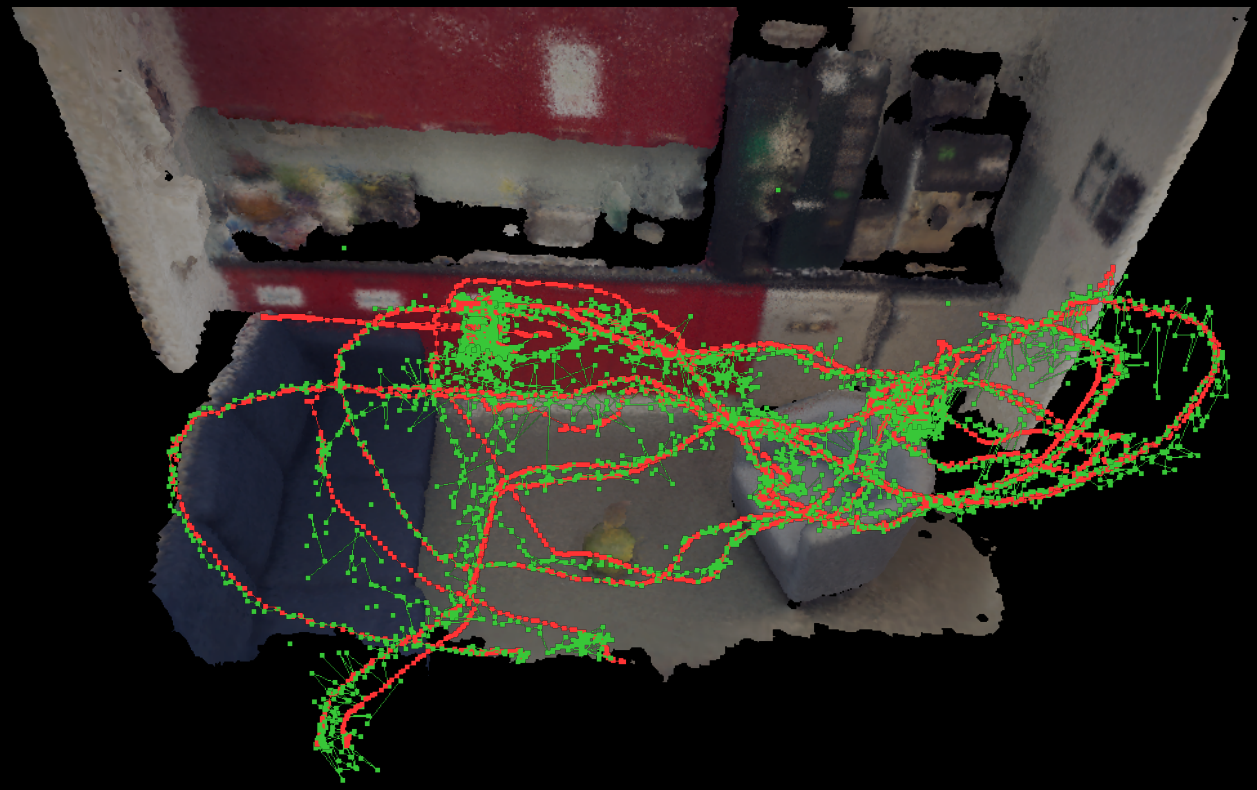

D-SLAM
(Kinect Fusion)
SfM
(COLMAP)



Active Search

D-SLAM (Kinect Fusion)
pseudo ground truth

T. Sattler, B. Leibe, and L. Kobbelt. “Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization”. PAMI, 2017



Active Search

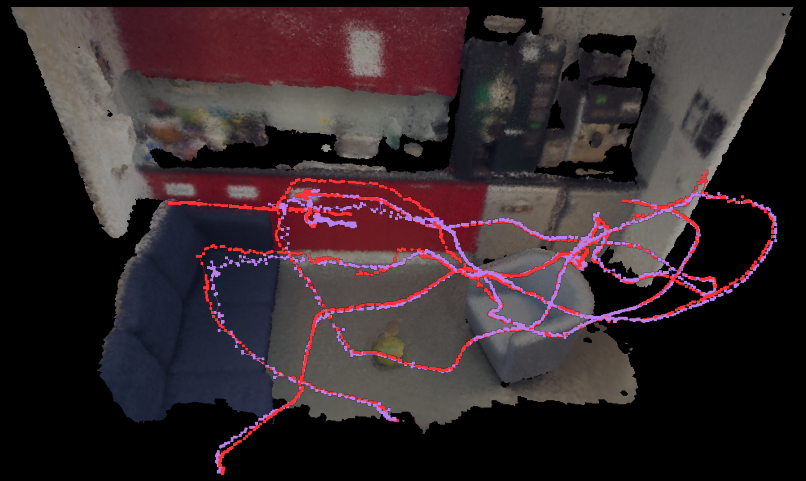
DSO (KeyFrame)

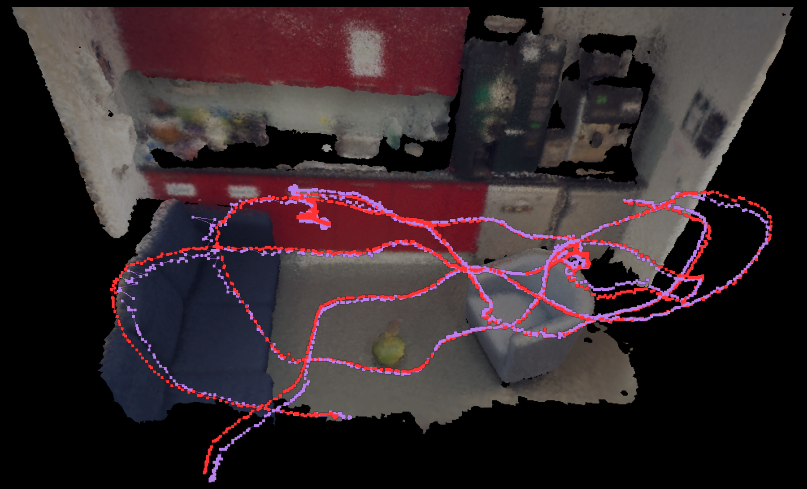
DSO (Point Fusion)

pseudo ground truth

T. Sattler, B. Leibe, and L. Kobbelt. “Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization”. PAMI, 2017

DSAC* (RGB-D)

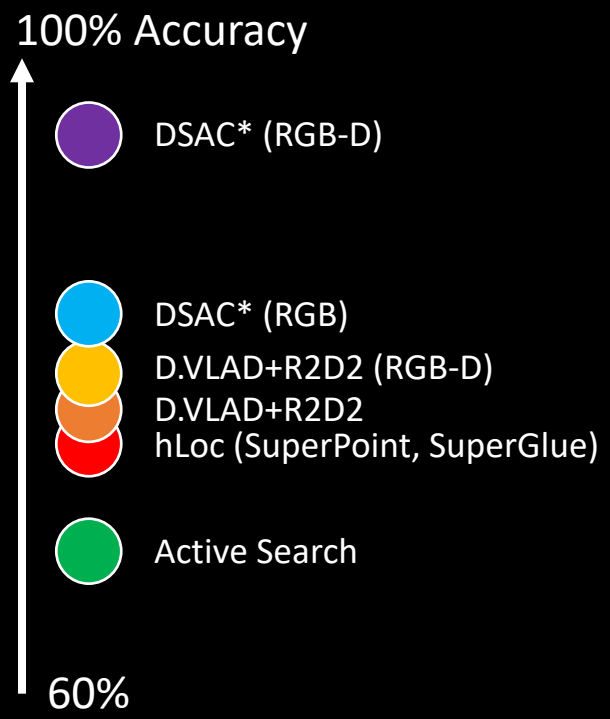




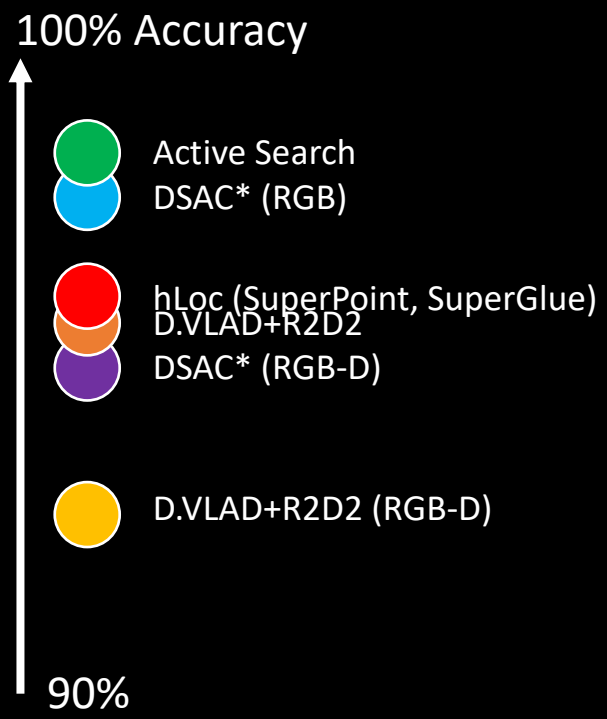
D-SLAM pseudo ground truth (Kinect Fusion)

SfM pseudo ground truth (COLMAP)

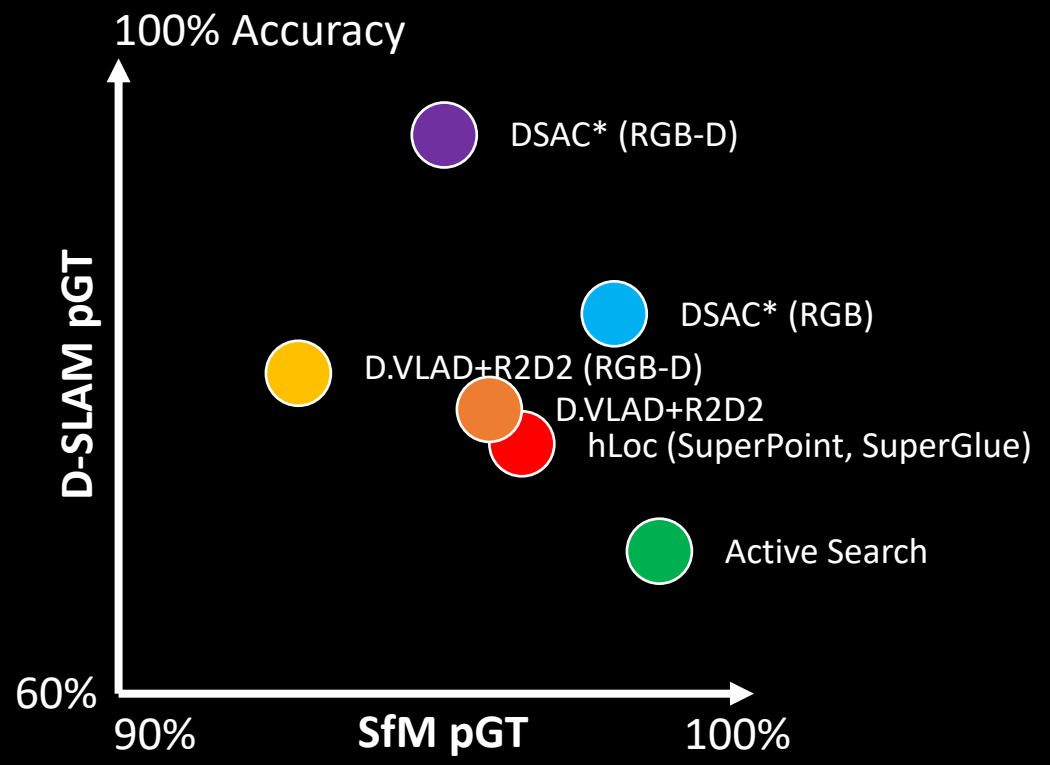
E. Brachmann and C. Rother. “Visual camera relocalization from RGB and RGB-D images using DSAC”. TPAMI, 2021



D-SLAM pGT (Kinect Fusion)



SfM pGT (COLMAP)



"Classic methods do not work well indoors."
"Scene coordinate regression outperforms classic re-localisation."
"RGB-D methods outperform RGB methods."

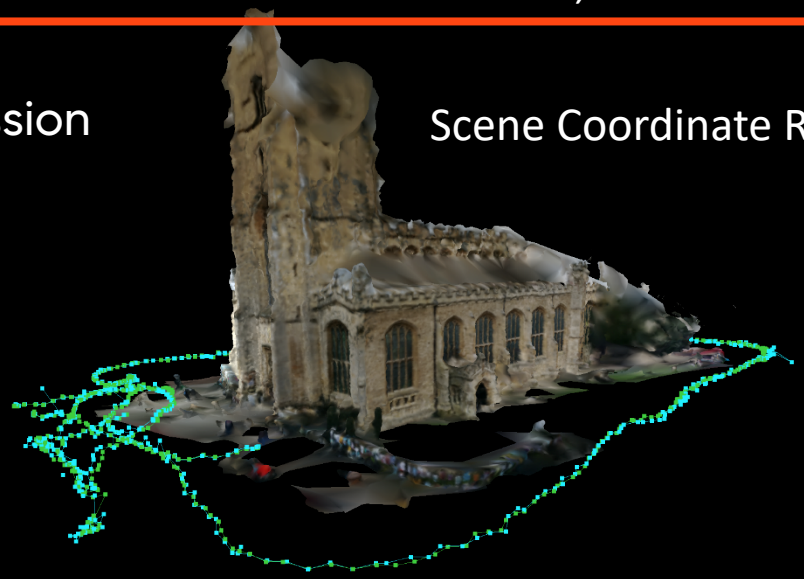
?

APR (The Challenge) Will there every be enough data?
RPR (The Promise) Fixing APR but not generalizing yet.
SCoRe (The Compromise) Keep domain knowledge, learn heuristics from data.

Houston: We have a benchmarking problem.

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

Absolute Pose Regression



Scene Coordinate Regression

DSAC* code: https://github.com/vislearn/dsacstar

Thank You!