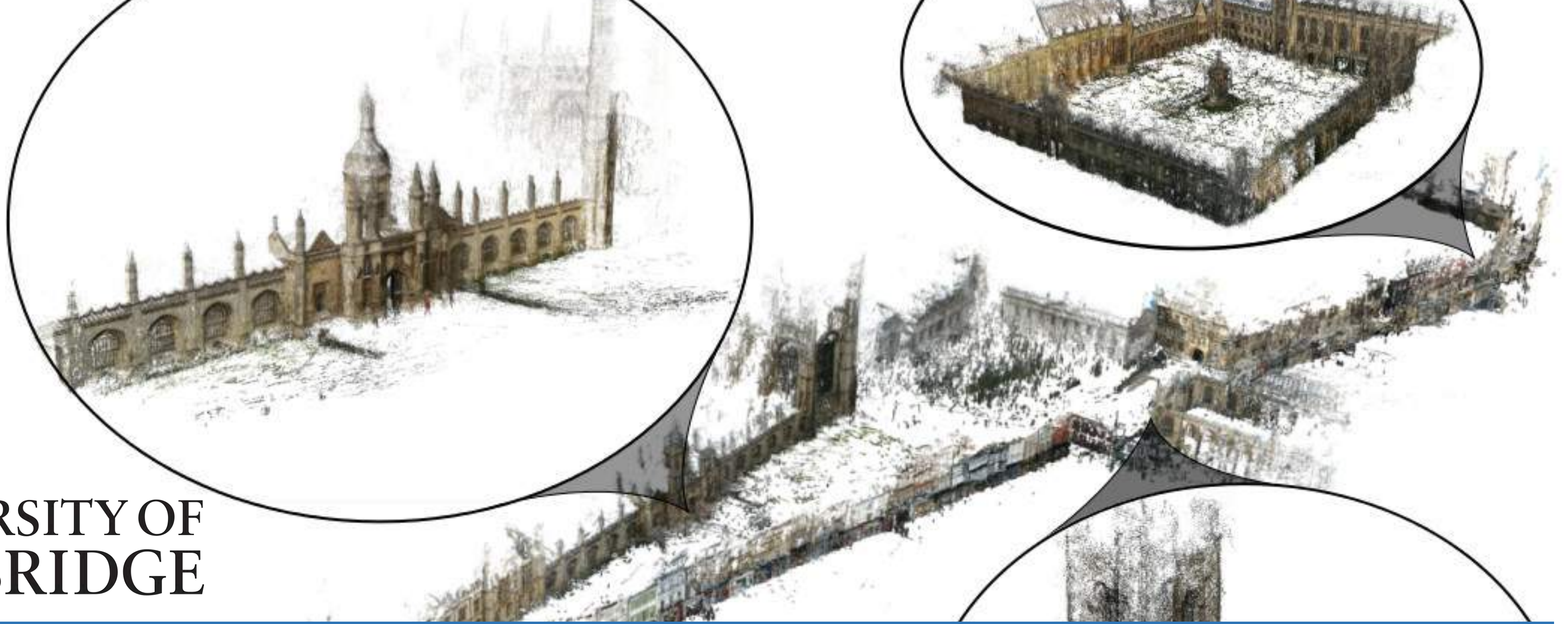


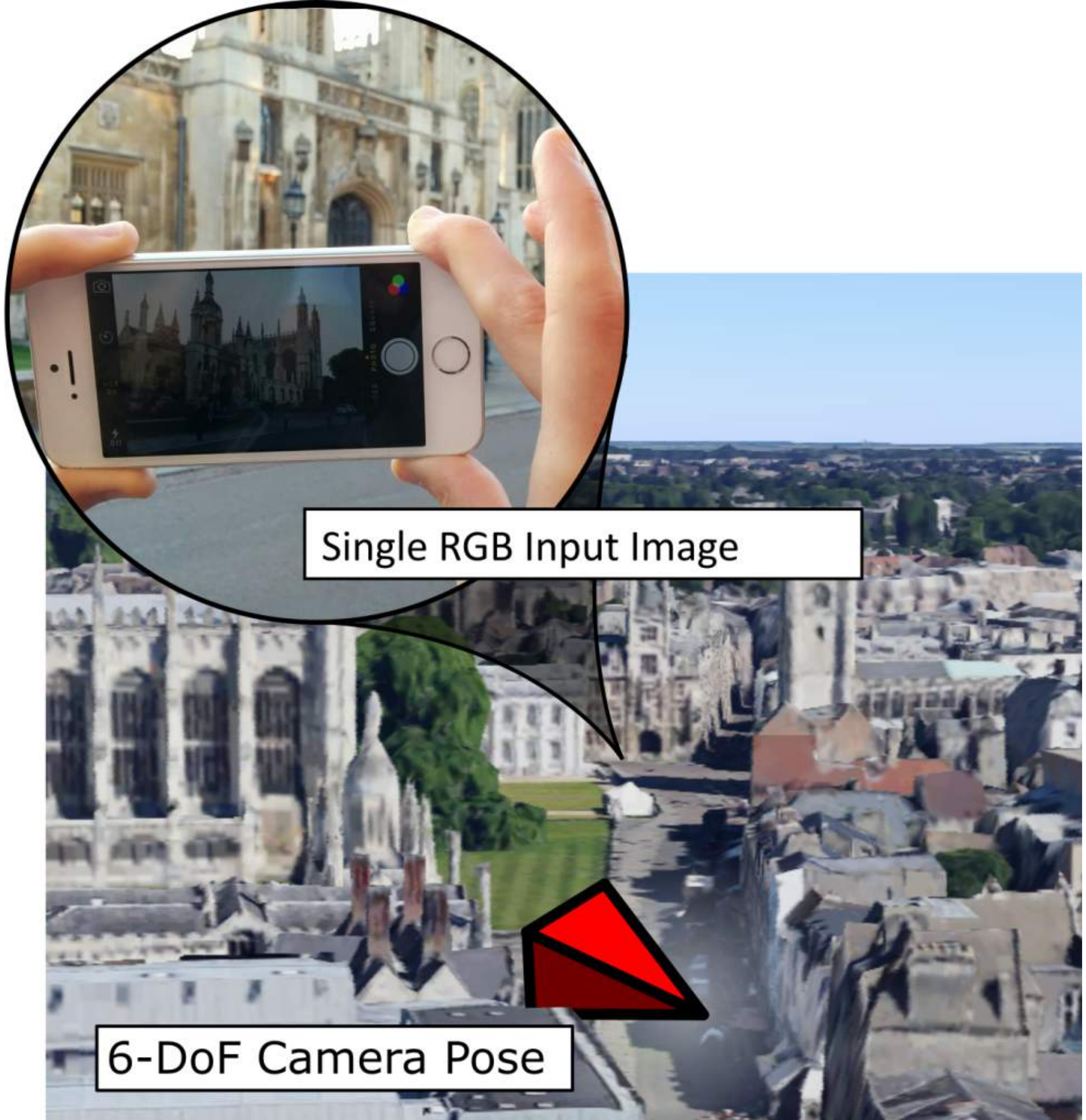
UNIVERSITY OF CAMBRIDGE

Learning-based Visual Localization

Alex Kendall, University of Cambridge

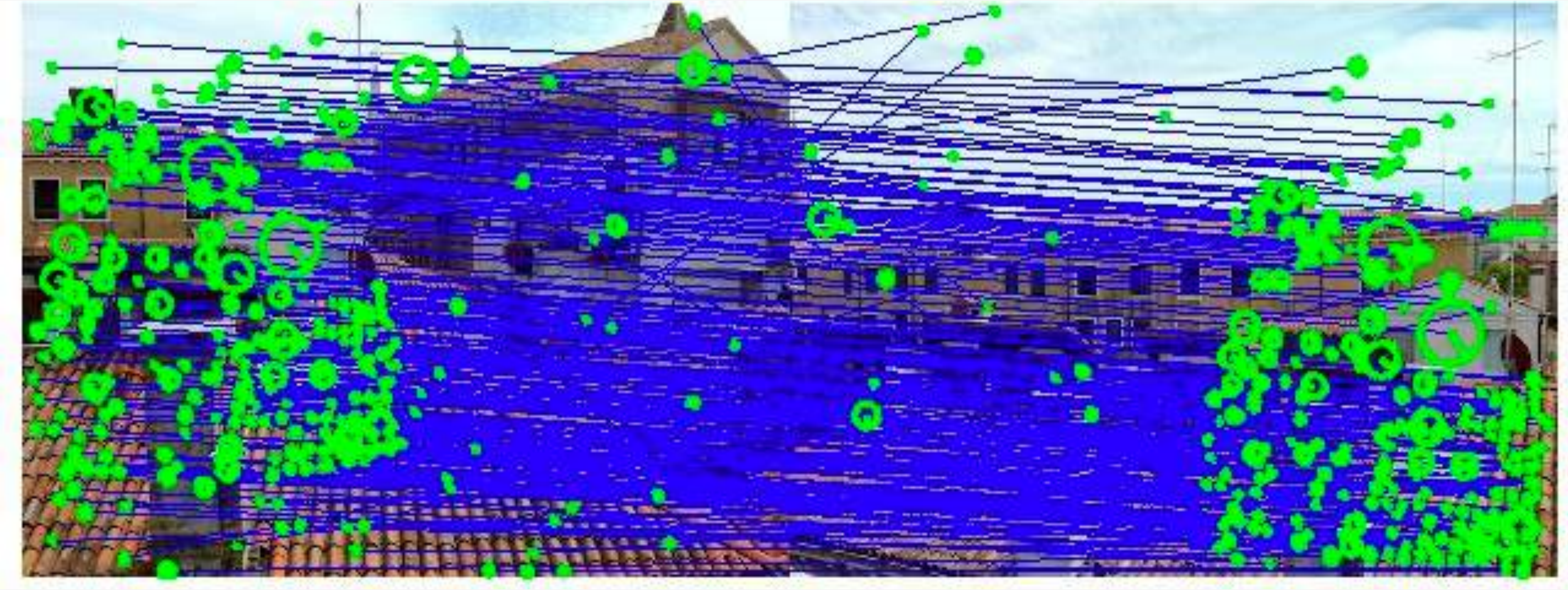
CVPR 2017 tutorial on Large-Scale Visual Place Recognition and Image-Based Localization

Alex Kendall, Torstel Sattler, Giorgos Tolias, Akihiko Torii


Single RGB Input Image
6-DoF Camera Pose

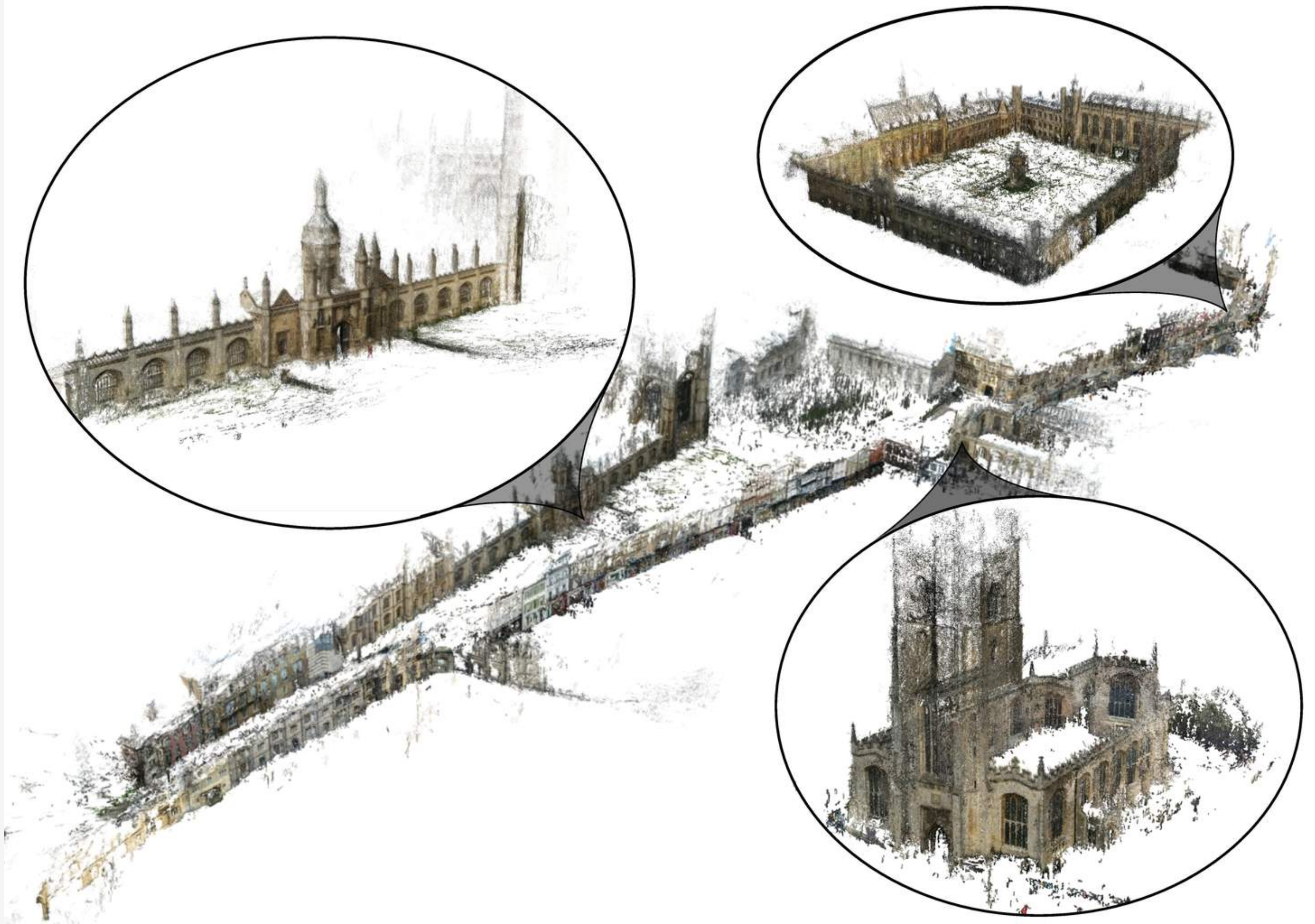
What is the motivation to use machine learning to localise?

Are point-based feature descriptors the right landmarks?



Also, should maps be Euclidean?

- Not all places are of equal importance?
- Can we learn a better 'map'?

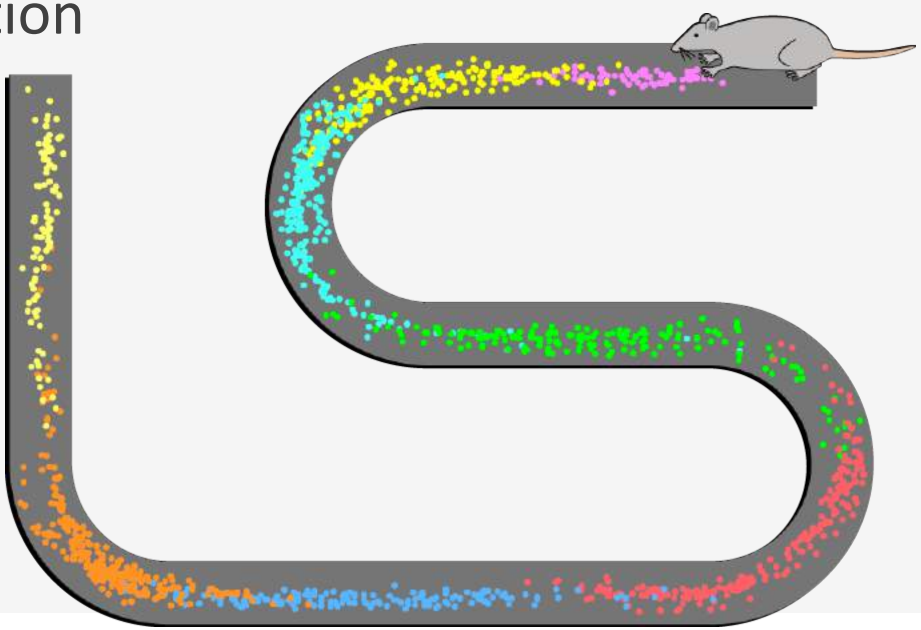


Biological learning for localisation

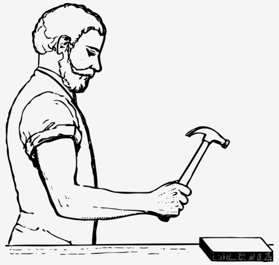
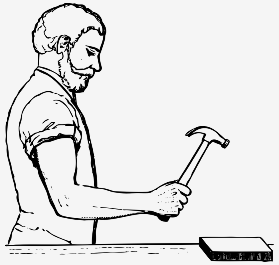
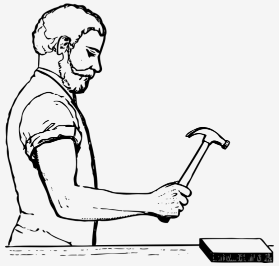
2014 Nobel Prize in Physiology or Medicine for the discovery of place and grid cells

[O'Keefe, Edvard and May-Britt Moser]

- Located in the hippocampus
- **Place cells** encode topological and hierarchical location
- **Grid cells** encode Euclidean space for precise positioning and path finding

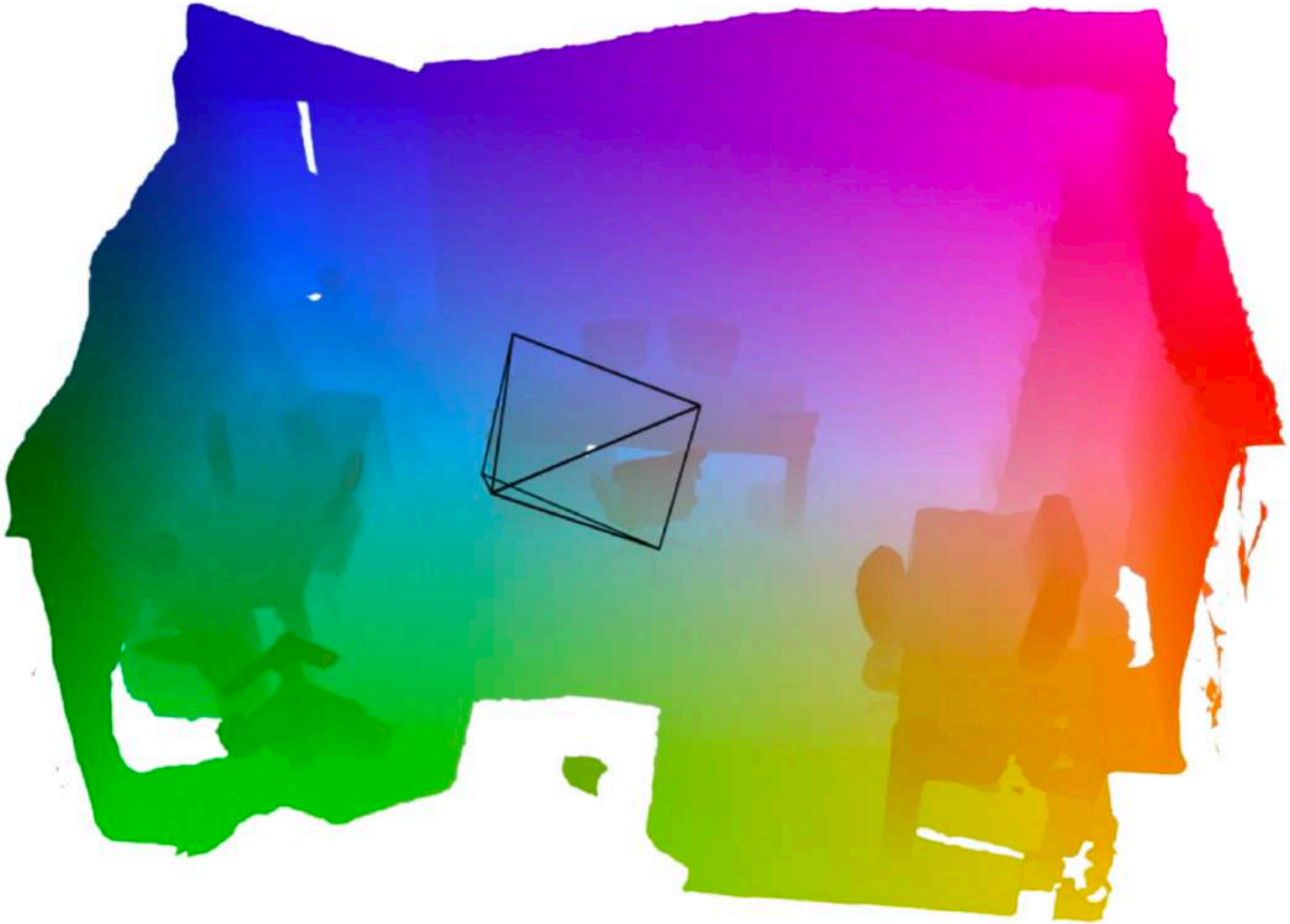


Machine Learning for Localisation

	Feature Extraction	Landmark Registration	6DOF Pose Estimation
Feature based localisation (previous talk)	Machine learning		
Scene Coordinate Regression (Part 1)	Machine learning	Machine learning	
Pose Regression (Part 2)	Machine learning	Machine learning	Machine learning

Part 1: Scene Coordinate Regression

Part 1: Scene Coordinate Regression



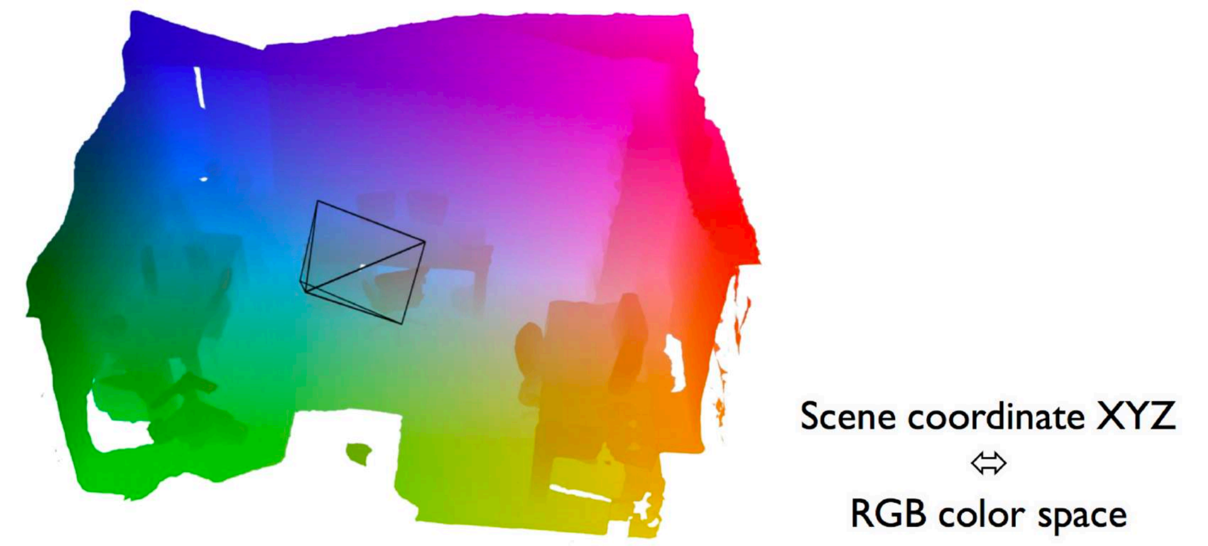
Scene coordinate XYZ

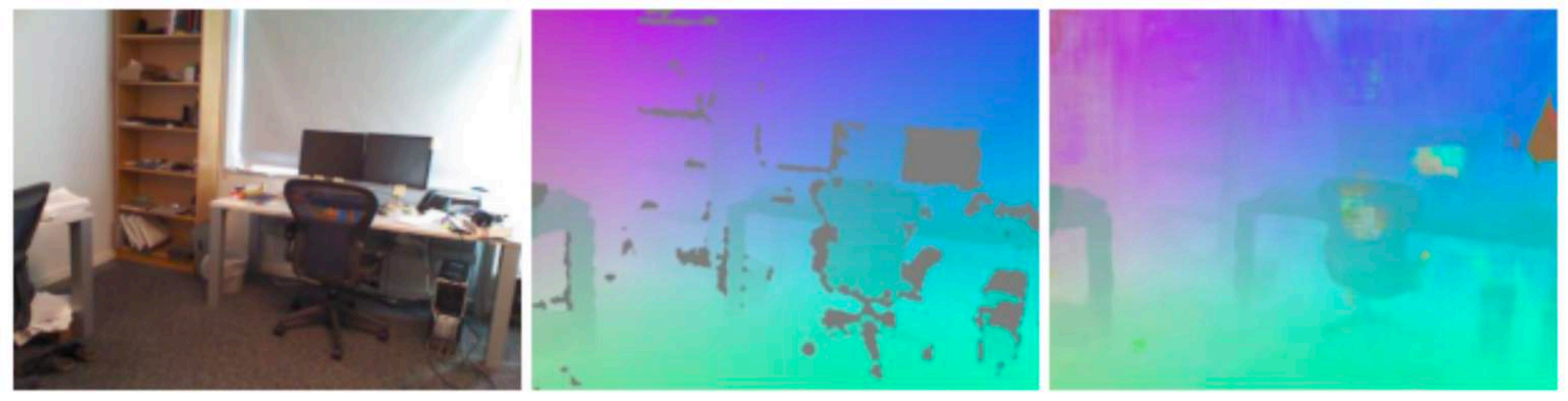
⇔

RGB color space

Scene Coordinate Localisation

- Infer scene coordinates for each pixel location in a test image



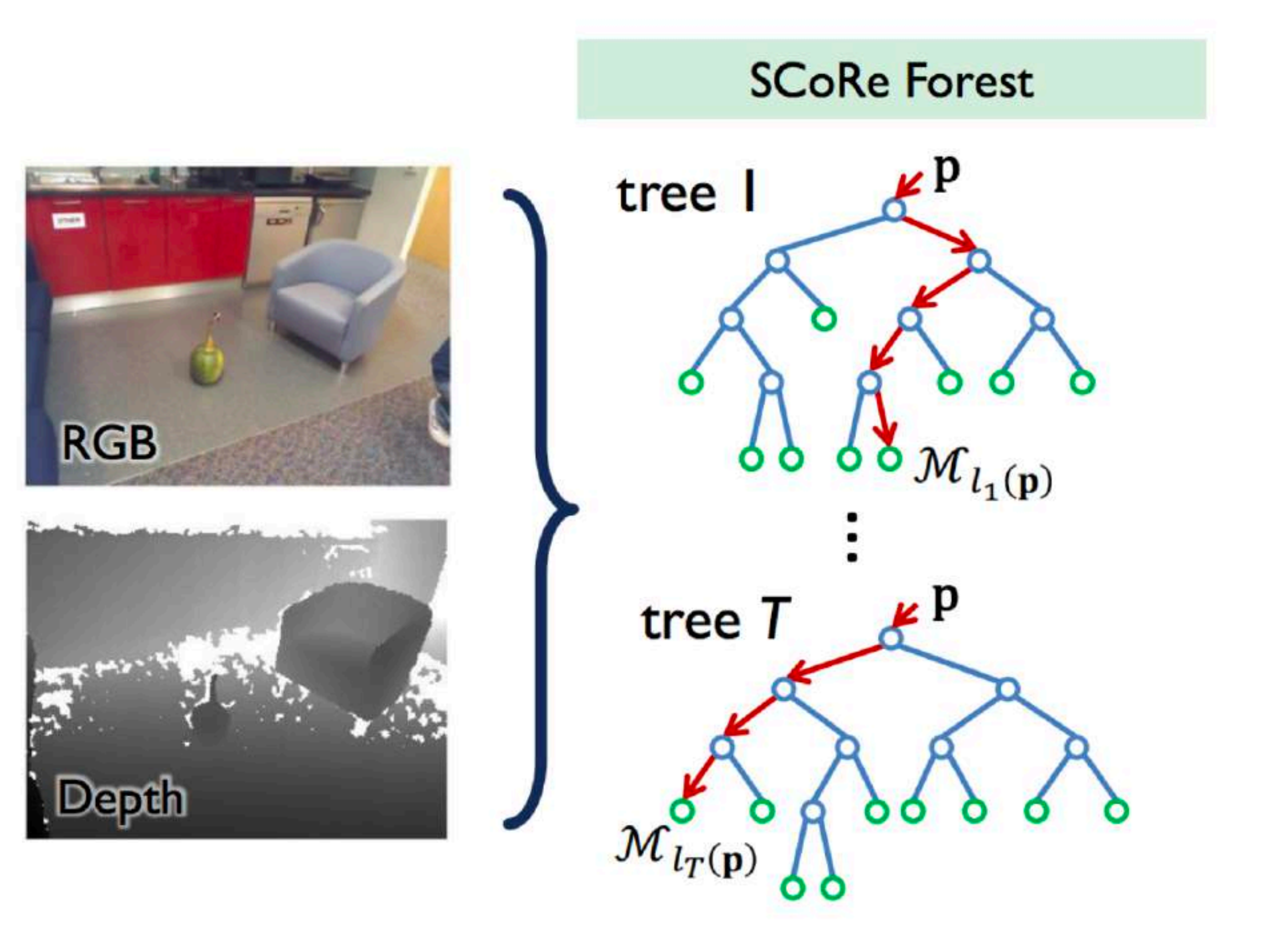


Input Image   Ground Truth Scene Coordinates   Predicted Scene Coordinates

Eric Brachmann et al. DSAC-Differentiable RANSAC for Camera Localization. *CVPR* 2017.

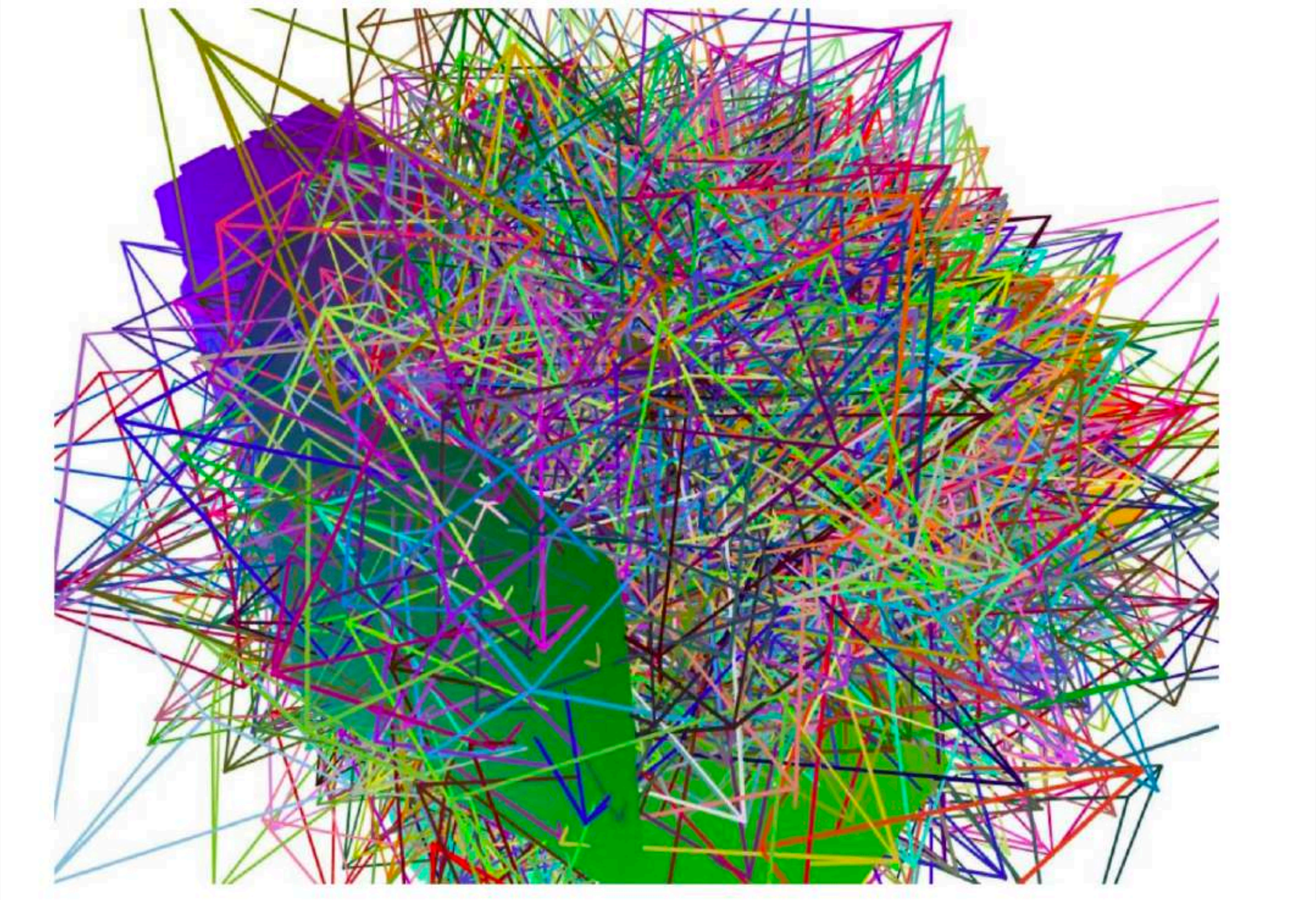
Scene Coordinate Regression

- Use SLAM/ Kinect Fusion / etc to generate ground truth labels
- Train a regression forest to regress scene coordinates
- Train on depth-aware features using RGB-D data



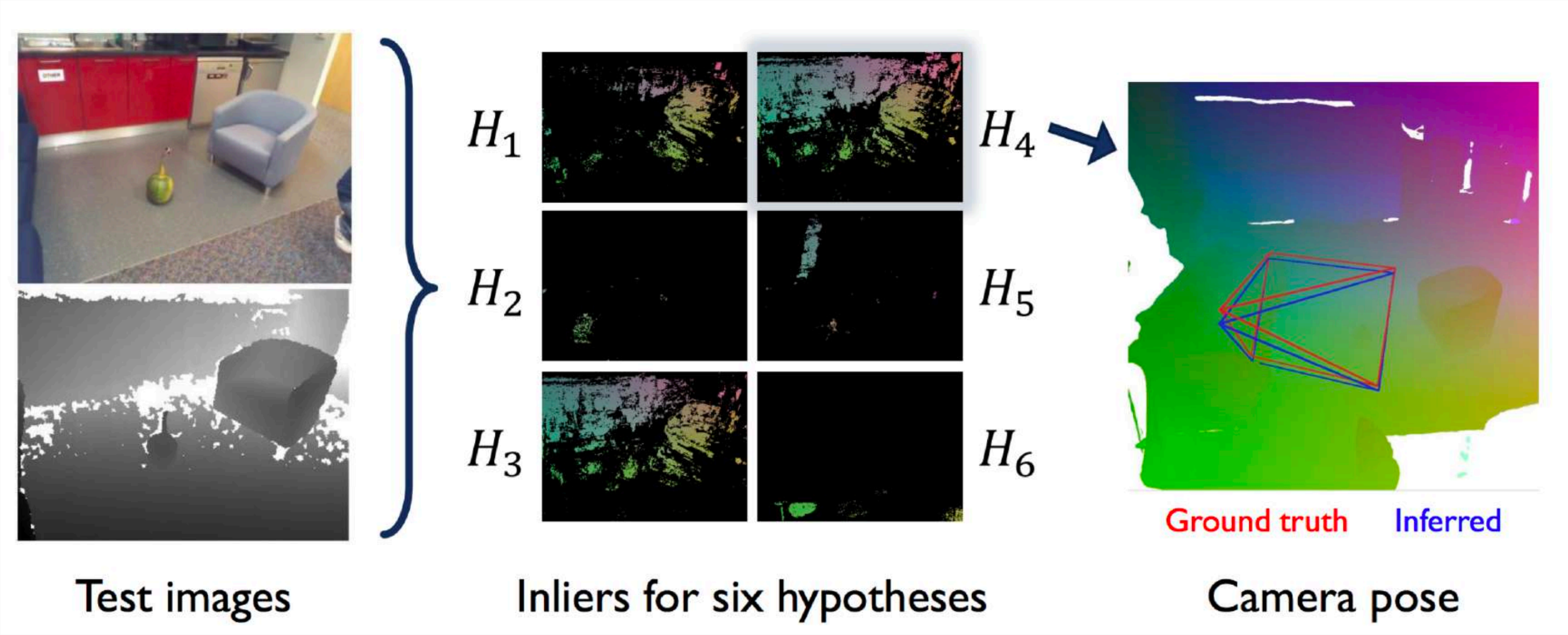
Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

Generate Camera Pose Hypothesis



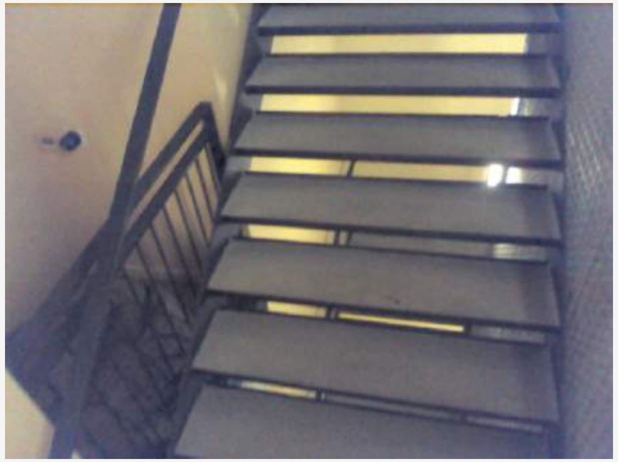
Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

RANSAC to infer camera pose



Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

Datasets – Seven Scenes – Indoor Localization




- 17,000 images across 7 small indoor scenes.

Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

Outperforms SIFT-Feature methods

Metric:

Proportion of test frames with < 0.05m translational error and < 5° angular error

Results:

Scene	Baselines: Tiny-image RGB-D	Baselines: Sparse RGB	Our Results: Depth	Our Results: DA-RGB	Our Results: DA-RGB + D
Chess	0.0%	70.7%	82.7%	**92.6%**	91.5%
Fire	0.5%	49.9%	44.7%	**82.9%**	74.7%
Heads	0.0%	**67.6%**	27.0%	49.4%	46.8%
Office	0.0%	36.6%	65.5%	74.9%	**79.1%**
Pumpkin	0.0%	21.3%	58.6%	**73.7%**	72.7%
RedKitchen	0.0%	29.8%	61.3%	71.8%	**72.9%**
Stairs	0.0%	9.2%	12.2%	**27.8%**	24.4%

SIFT feature registration

Different features for regression forest

Further Literature

Daniela Massiceti et al. Random Forests versus Neural Networks-What's Best for Camera Relocalization? *arXiv* 2016.

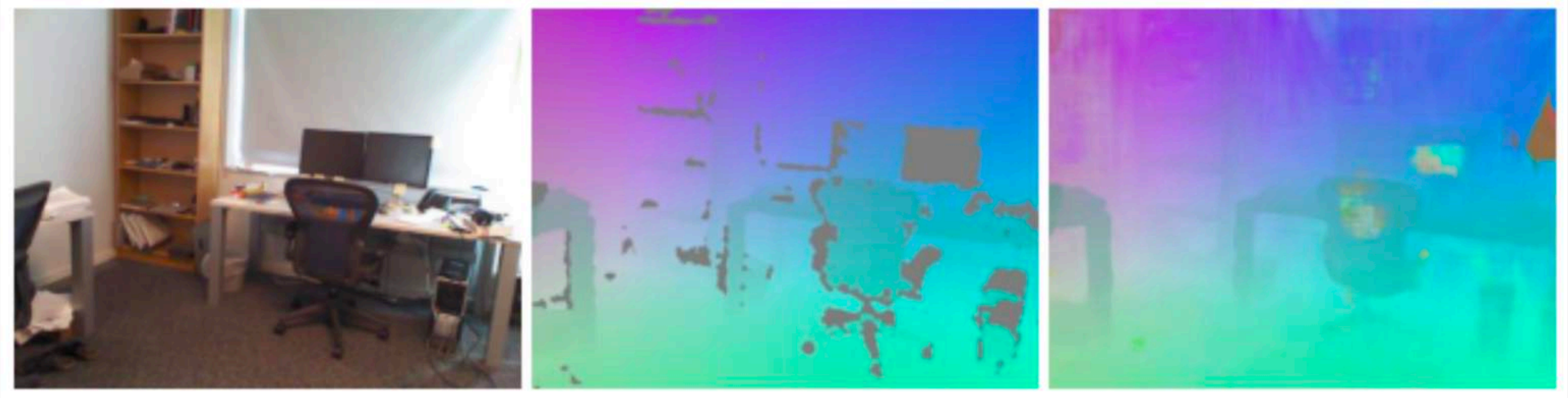
✓ Neural networks improve scene coordinate regression but not for RANSAC optimization

Eric Brachmann et al. Uncertainty-driven 6d pose estimation of objects and scenes from a single RGB image. CVPR 2016.

✓ Scene coordinate regression with RGB only images

Eric Brachmann et al. DSAC-Differentiable RANSAC for Camera Localization. *CVPR* 2017.

✓ End-to-end learning through RANSAC with deep learning



Input Image   Ground Truth Scene Coordinates   Predicted Scene Coordinates

Score Regression Conclusions

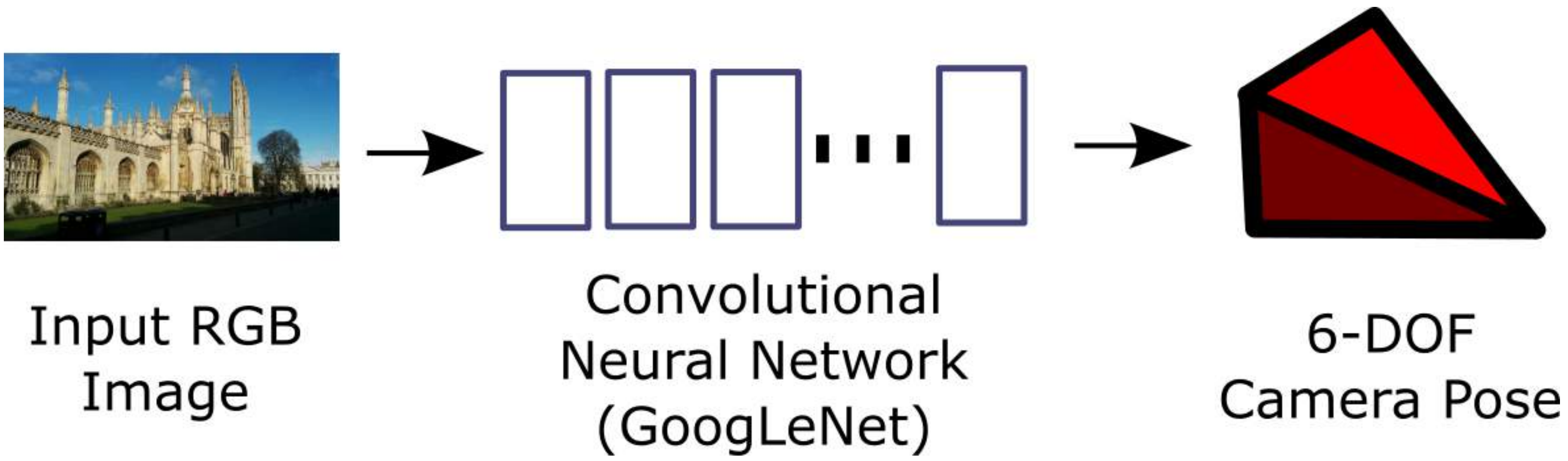
✓Map no longer Euclidean but learned features

✓Scales efficiently with scene size

✗Features don’t use global context

✗Inliers are noisy and RANSAC not always reliable

Part 2: Pose Regression with Deep Learning



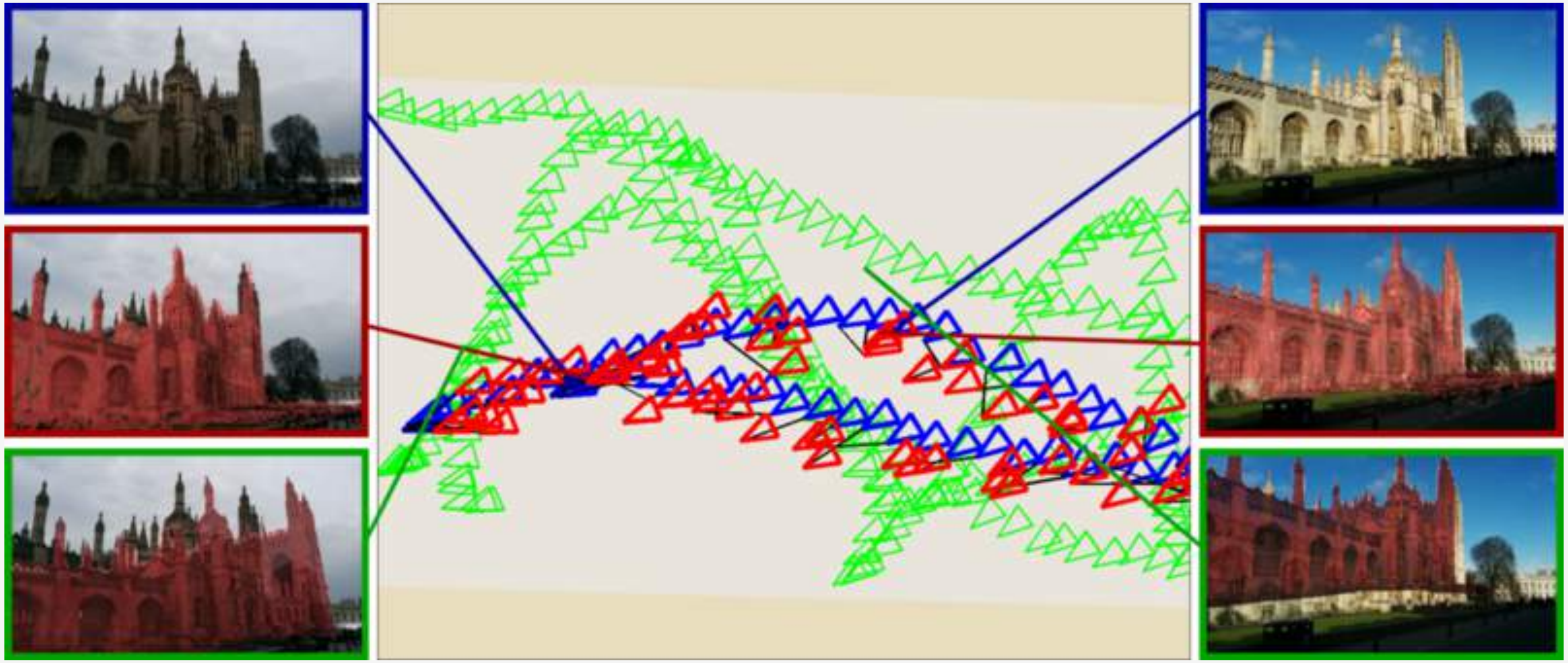
Trained with a naïve end-to-end loss function to regress camera position, **x**, and orientation, **q**

$$\mathrm{loss}(I) = \|x - \hat{x}\|_2 + \beta \left\| q - \frac{\hat{q}}{\|\hat{q}\|} \right\|_2$$

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

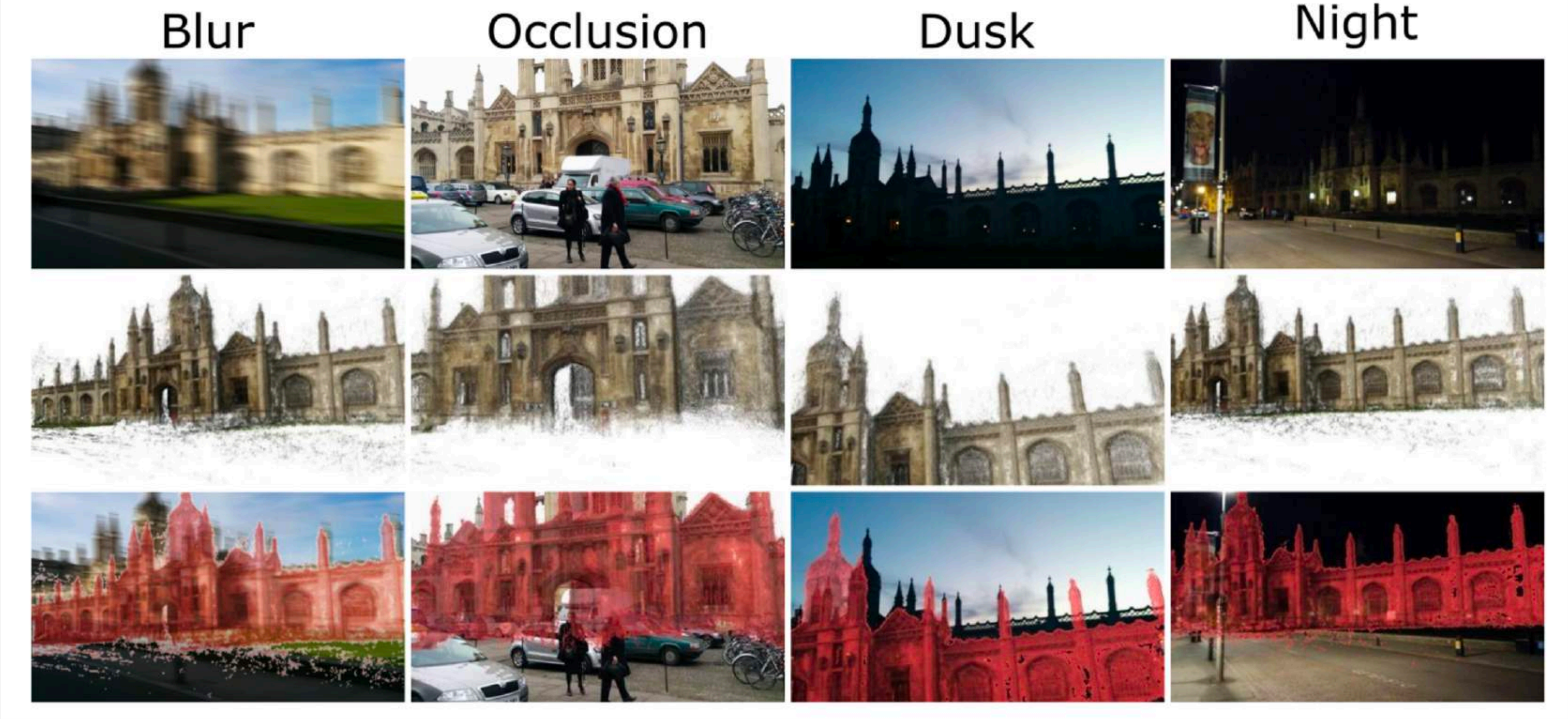
Camera Pose Regression

training data in green, test data in blue, PoseNet results in red



Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

Tolerance to environment, unknown intrinsics, weather, etc.



Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

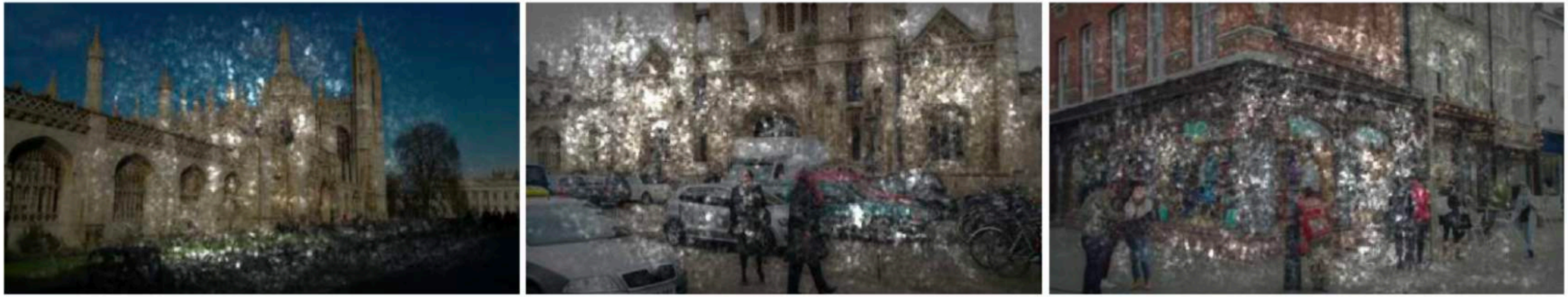
Robust in scenarios where SIFT-Feature localisation fails

SIFT based registration fails, but deep learning features are able to localise!

Area	# train/test	PoseNet [23]	Proposed
5575 m^2	875/220	1.87 m, 6.14°	1.31 m, 2.79° (30,55)

Florian Walch, et al. "Image-based localization using LSTMs for structured feature correlation." arXiv:1611.07890. 2016.

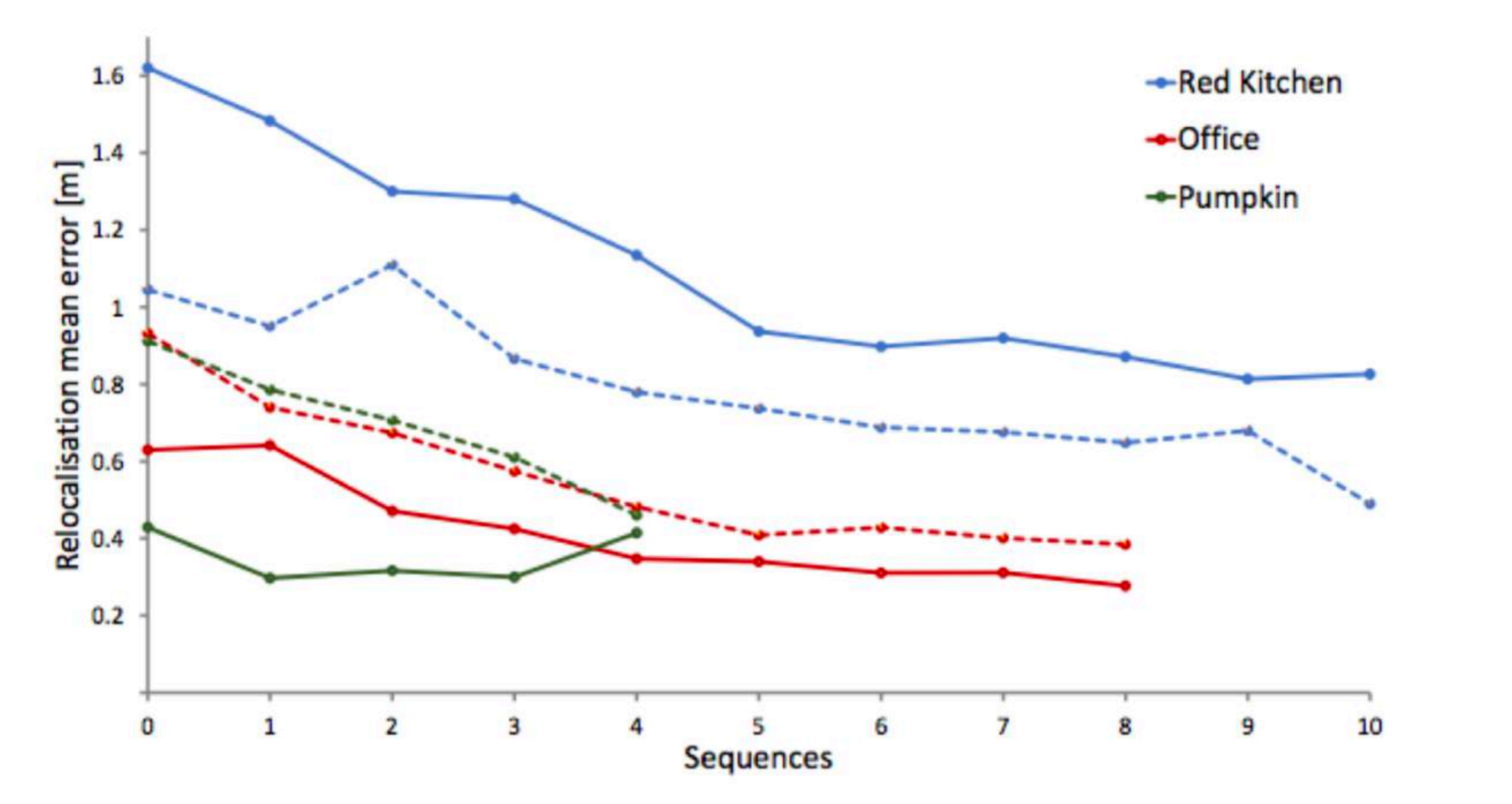
Does our network recognise context?



- Saliency maps show significance of each pixel w.r.t. localisation
- PoseNet learns to ignore dynamic objects and recognises large context and contours around landmarks

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

PoseNet's performance improves with more data



Scales very well:

- Constant inference time (single forward pass of the network)
- Constant memory (~5 MB of neural network weights)

Contreras, Luis, and Walterio Mayol-Cuevas. **Towards CNN Map Compression for camera relocalisation.** arXiv:1703.00845, 2017.

PoseNet Summary

✓Robust to lighting, weather, dynamic objects

✓Fast inference, <2ms per image on Titan GPU

✓Scale not dependent on number of training images

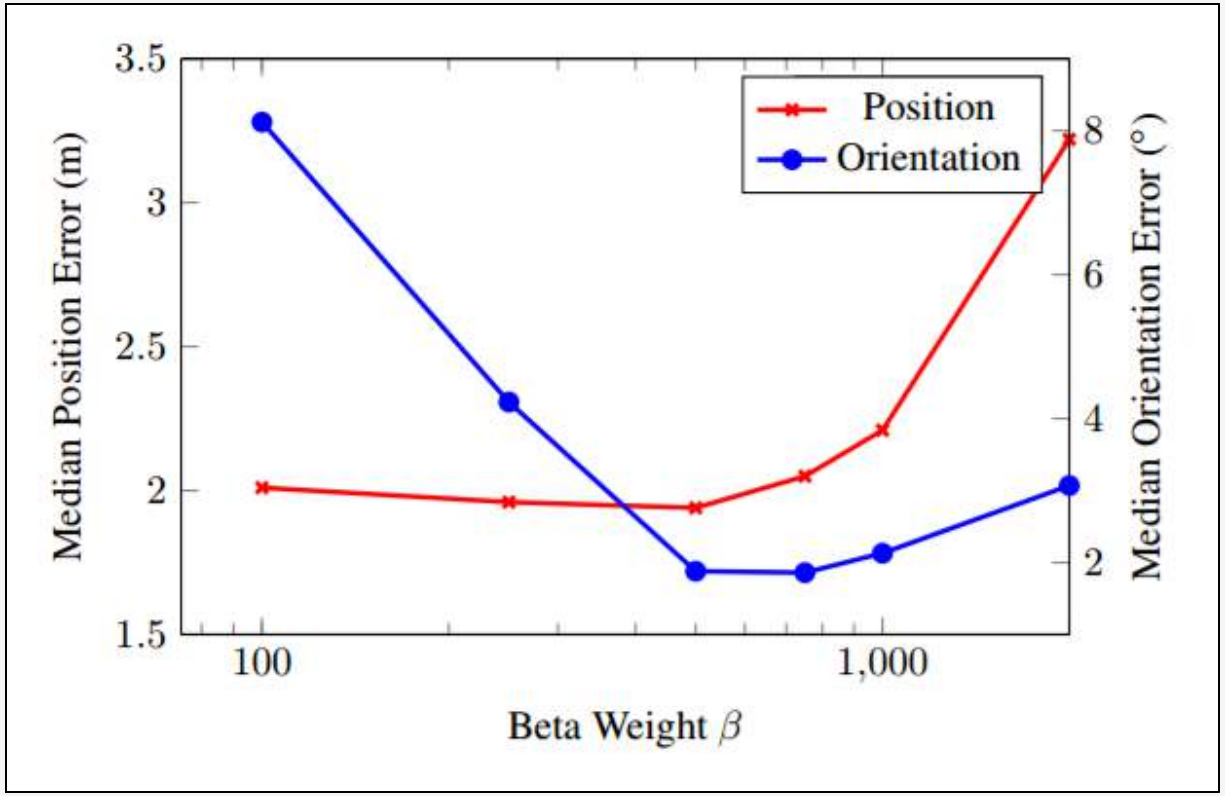
✗Coarse accuracy

✗Difficult to learn both position vs orientation

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

Intolerant to weighting between position and orientation regression loss

$$\mathrm{loss}(I) = \|x - \hat{x}\|_2 + \beta \left\| q - \frac{\hat{q}}{\|\hat{q}\|} \right\|_2$$

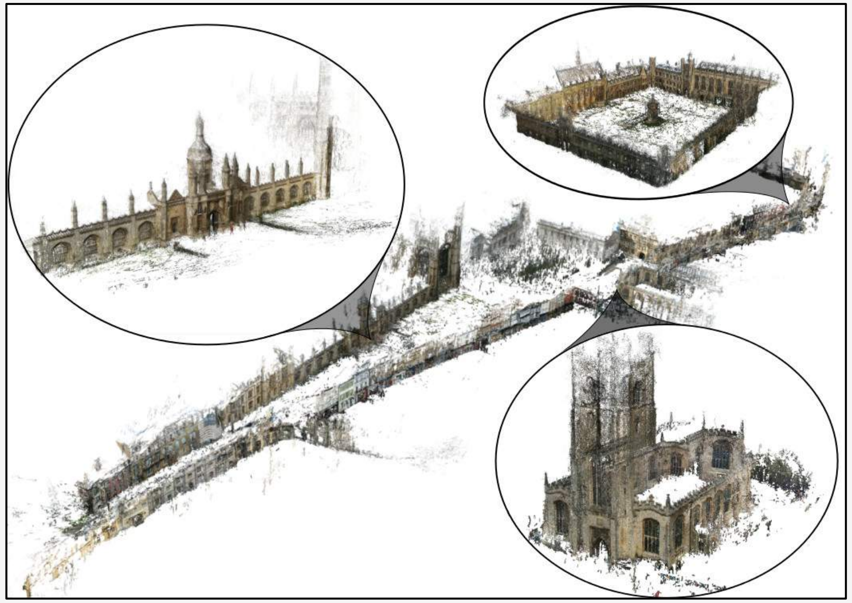


Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

Learning camera pose, *with geometry*

Train with reprojection loss of 3-D geometry using predicted and ground truth camera poses.

$$loss(I) = \frac{1}{|\mathcal{G}'|} \sum_{g_i \in \mathcal{G}'} \left\| \pi(\mathbf{q}, \mathbf{x}, \mathbf{g_i}) - \pi(\hat{\mathbf{q}}, \hat{\mathbf{x}}, \mathbf{g_i}) \right\|_\gamma$$



Where $\boldsymbol{\pi}$ is the projection function of 3-D point $\boldsymbol{g_i}$

Automatically learns a weighting between position, **x**, and orientation, **q**!

Alex Kendall and Roberto Cipolla. **Geometric loss functions for camera pose regression with deep learning.** CVPR, 2017.

Datasets – Cambridge Landmarks – Outdoor Localization

- 8,000 images from 6 scenes up to 100 x 500m

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

Datasets – Seven Scenes – Indoor Localization

- 17,000 images across 7 small indoor scenes.

Jamie Shotton et al. Scene coordinate regression forests for camera relocalization in RGB-D images. CVPR 2013

Datasets – Dubrovnik – Large Scale Localization










- 6000 images across 1500 x 1500 m in Dubrovnik, Croatia.
- Varying weather, season, camera type

Li, Yunpeng, Noah Snavely, and Daniel P. Huttenlocher. "Location recognition using prioritized feature matching." ECCV, 2010.

Geometry Improves Performance

	Cambridge Landmarks, King's College			Dubrovnik 6K		
	Median Error		**Accuracy**	**Median Error**		**Accuracy**
Loss function	x[m]	q[◦]	< 2m,5◦	x[m]	q[◦]	< 5m,5◦
Linear sum, β = 500 [1]	1.52	1.19	65%	13.1	4.68	30.1%
Learn weighting with task uncertainty [2]	0.99	1.06	85.3%	9.88	4.73	41.7%
Reprojection loss [2]	*does not converge*					
Learn weighting pretrain + Reprojection loss [2]	0.88	1.04	90.3%	7.90	4.40	48.6%
SIFT + SfM Geometry [3]	0.42	0.55	-	1.1	-	-

[1]. Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

[2]. Alex Kendall and Roberto Cipolla. **Geometric loss functions for camera pose regression with deep learning.** CVPR, 2017.

[3]. T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. PAMI, 2016.

Future Work & What's Next?

- PoseNet is much faster and requires smaller images than traditional methods

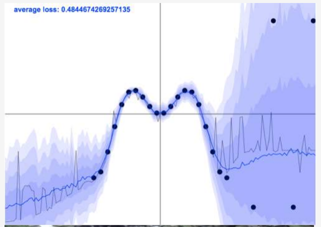
Dataset	PoseNet with Geometry [1]	Active Search (SIFT + Geometry) [2]
King's College	0.88m, 1.04°	**0.42m, 0.55°**
Resolution	**256 x 256 px**	1920 × 1080 px
Inference Time	**2 ms**	78 ms

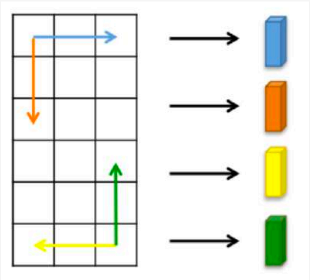
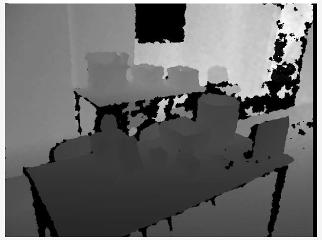
- Can we scale model towards city scale localisation with deep learning?
- How to improve fine grained accuracy for accurate registration?

[1]. Alex Kendall and Roberto Cipolla. **Geometric loss functions for camera pose regression with deep learning.** CVPR, 2017.

[2]. T. Sattler, B. Leibe, and L. Kobbelt. **Efficient & effective prioritized matching for large-scale image-based localization.** PAMI, 2016.

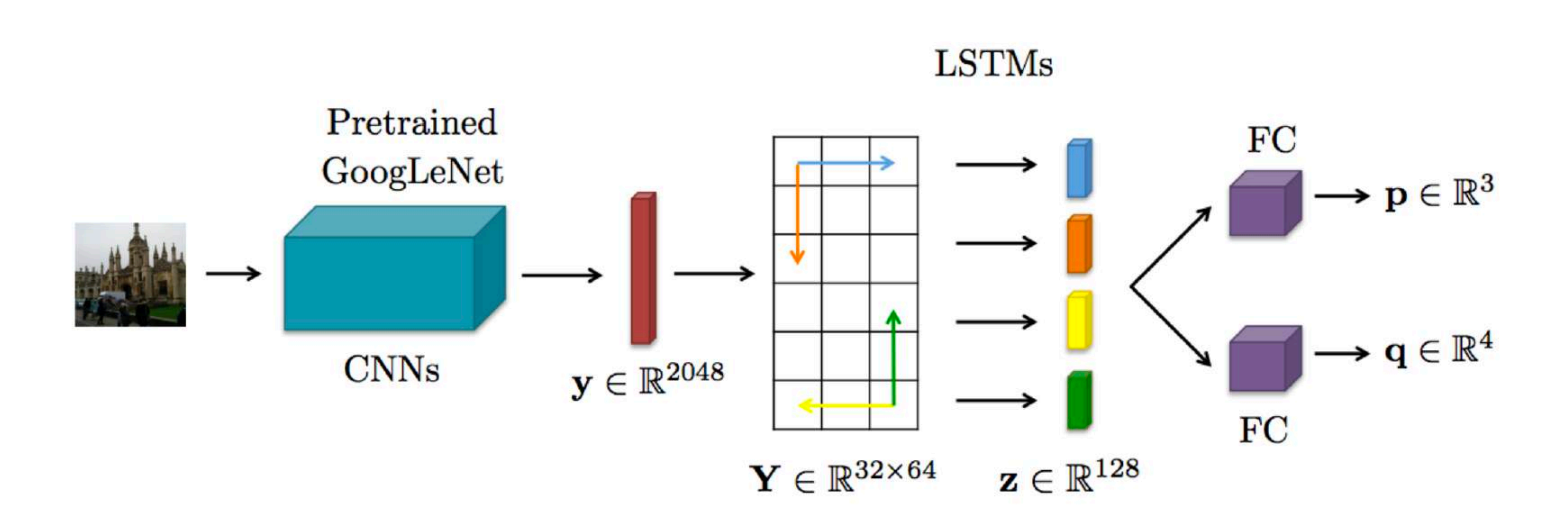
Further Improvements to Camera Pose Regression



- Modelling uncertainty: Kendall et al. ICRA 2017
- Map compression: Contreras et al. arXiv 2016
- Improve context of features: Walch et al. arXiv 2016
- Video localisation: Clark et al. CVPR 2017
- Ego-motion estimation: Melekhov et al. arXiv 2017
- RGB-D localisation: Li et al. IEEE Transactions 2017

Increase feature context with spatial LSTMs



Improves metric localisation performance from PoseNet by 5-50% (depending on the dataset)

Florian Walch et al. **Image-based localization using LSTMs for structured feature correlation.** arXiv 2016.

Modeling Aleatoric Uncertainty with Probabilistic Deep Learning

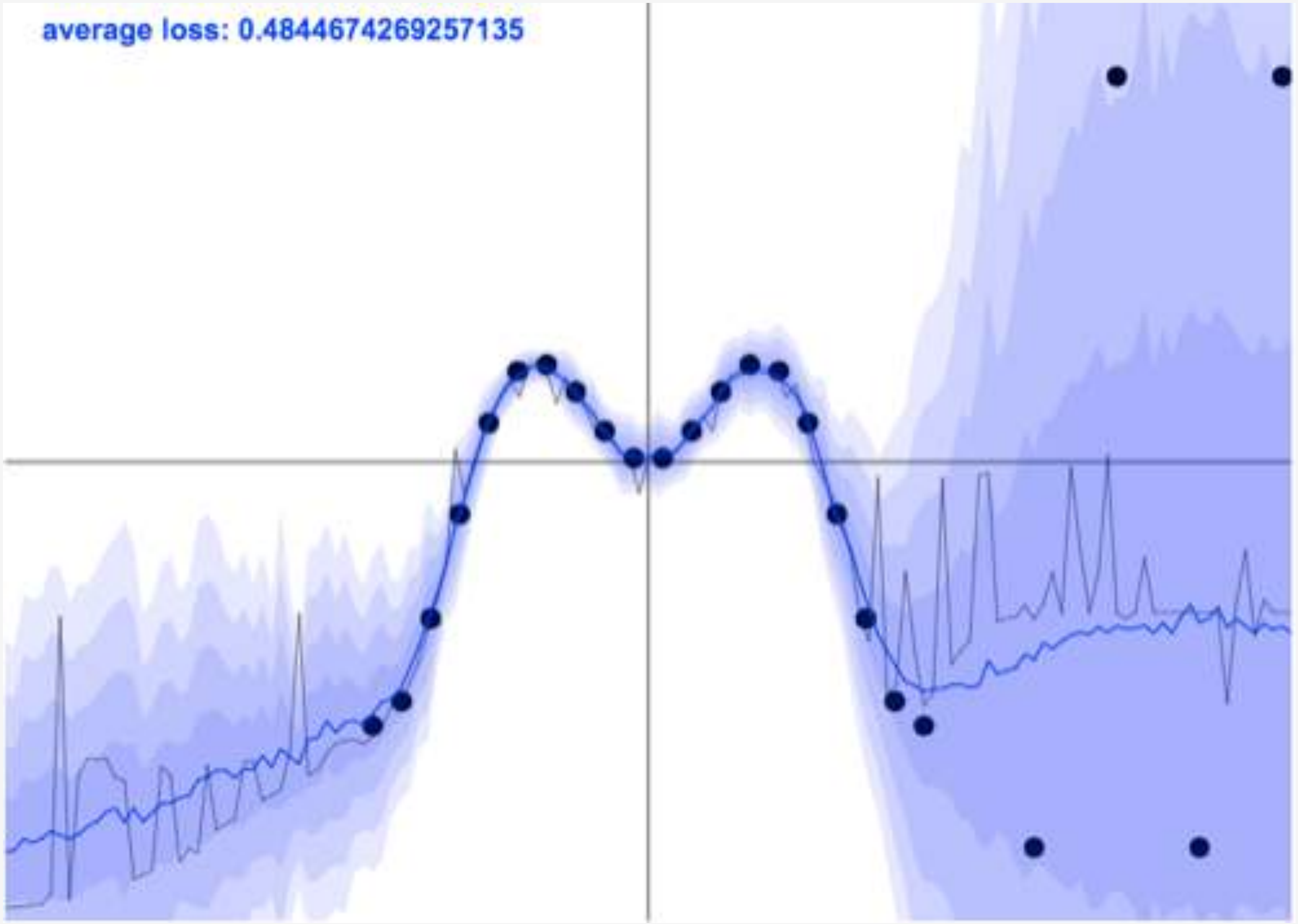
Use probabilistic modelling to learn data-dependant uncertainty with no additional supervision.

	Deep Learning	Probabilistic Deep Learning
Model	$[\hat{y}\,] = f(x)$	$[\hat{y}, \hat{\sigma}^2] = f(x)$
Regression	$Loss = \lVert y - \hat{y} \rVert^2$	$Loss = \frac{\lVert y - \hat{y} \rVert^2}{2\hat{\sigma}^2} + \log \hat{\sigma}^2$

Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** arXiv preprint 1703.04977, 2017.

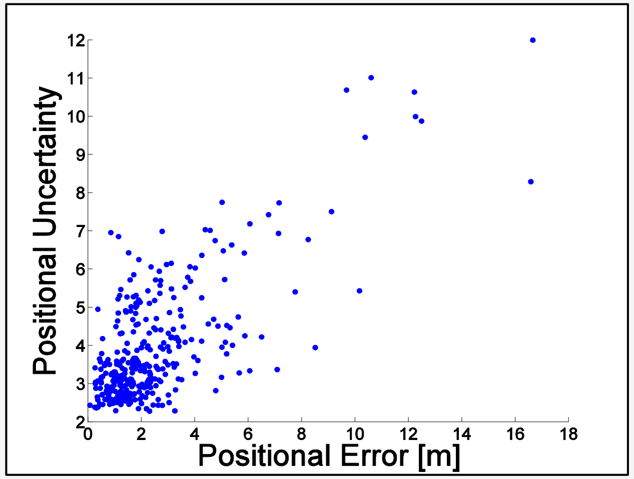
Model Uncertainty with Dropout

- Use dropout to learn a distribution over models
- Sample using Monte Carlo dropout sampling a test time to obtain posterior distribution

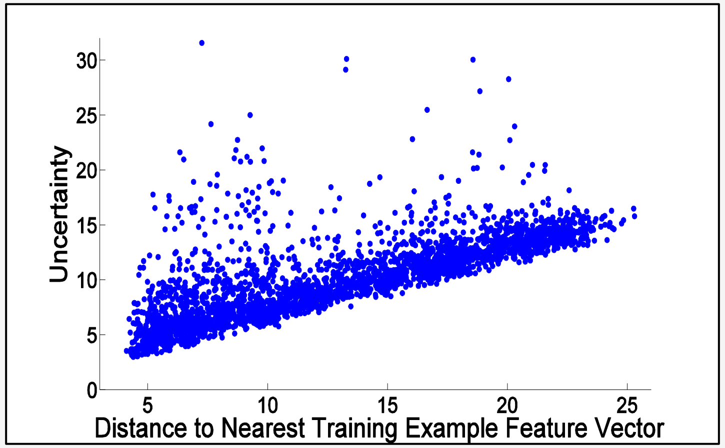


Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** arXiv preprint 1703.04977, 2017.

Uncertainty to estimate loop closure



We can use epistemic uncertainty to estimate metric relocalisation error



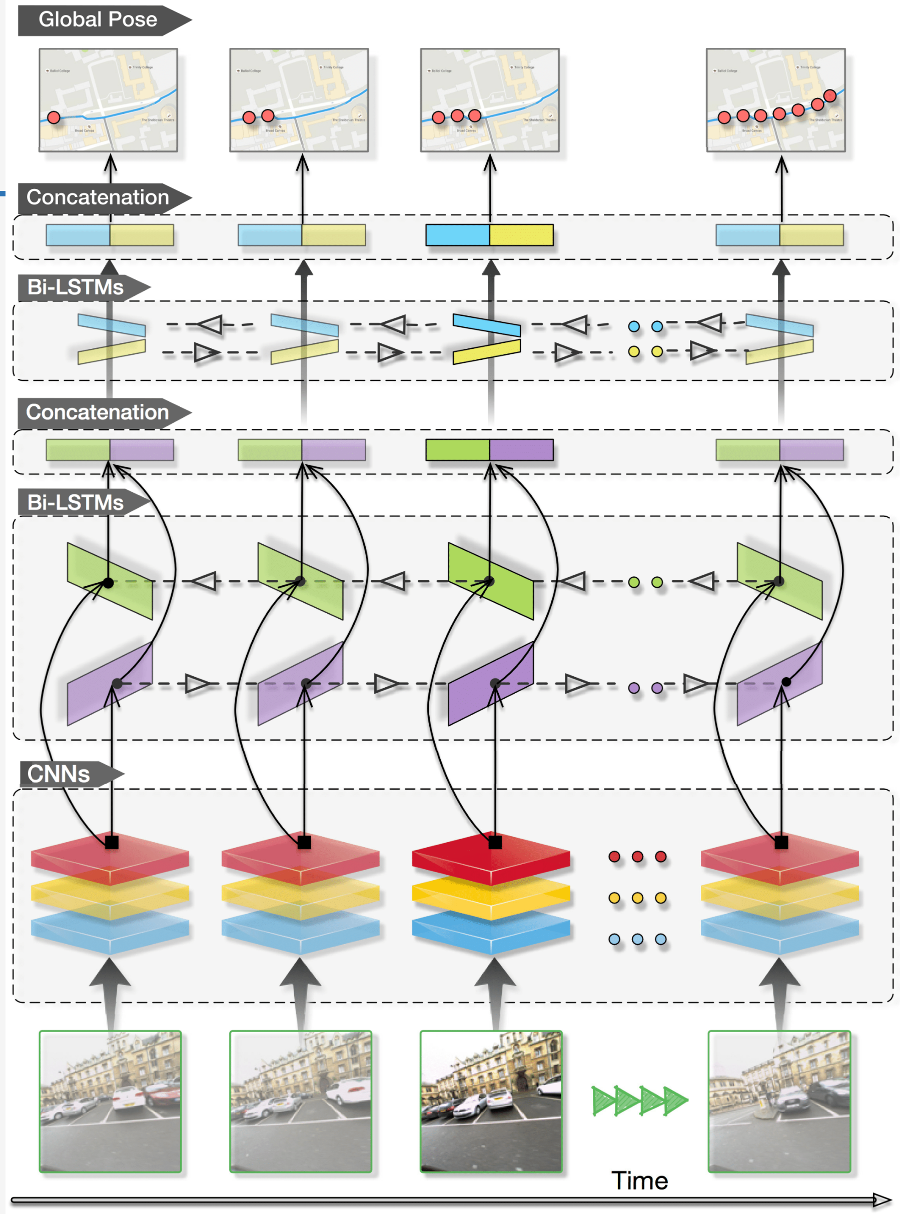
Determine if the model has seen the landmark before (loop closure)

Increased uncertainty from strong occlusion, motion blur, visually ambiguous landmarks

Alex Kendall and Roberto Cipolla. **Modelling Uncertainty in Deep Learning for Camera Relocalization.** ICRA, 2016.

Video Localisation

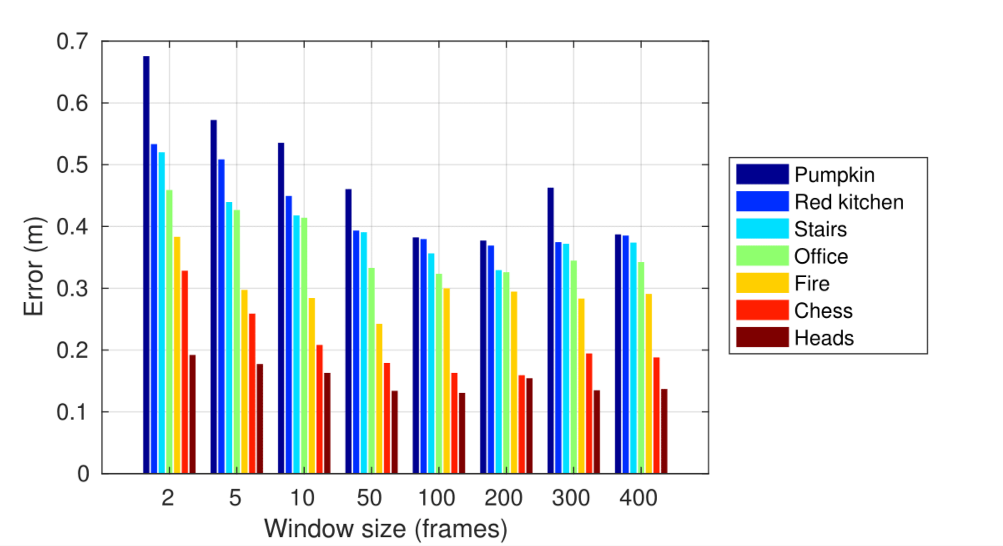
- PoseNet + Temporal Recurrent Neural Network
  - Learns dynamics of platform – temporal features
  - Bidirectional – analogous to "smoothing"
- Mixture of Gaussian output



Clark et al., ***VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization***. IEEE CVPR 2017.

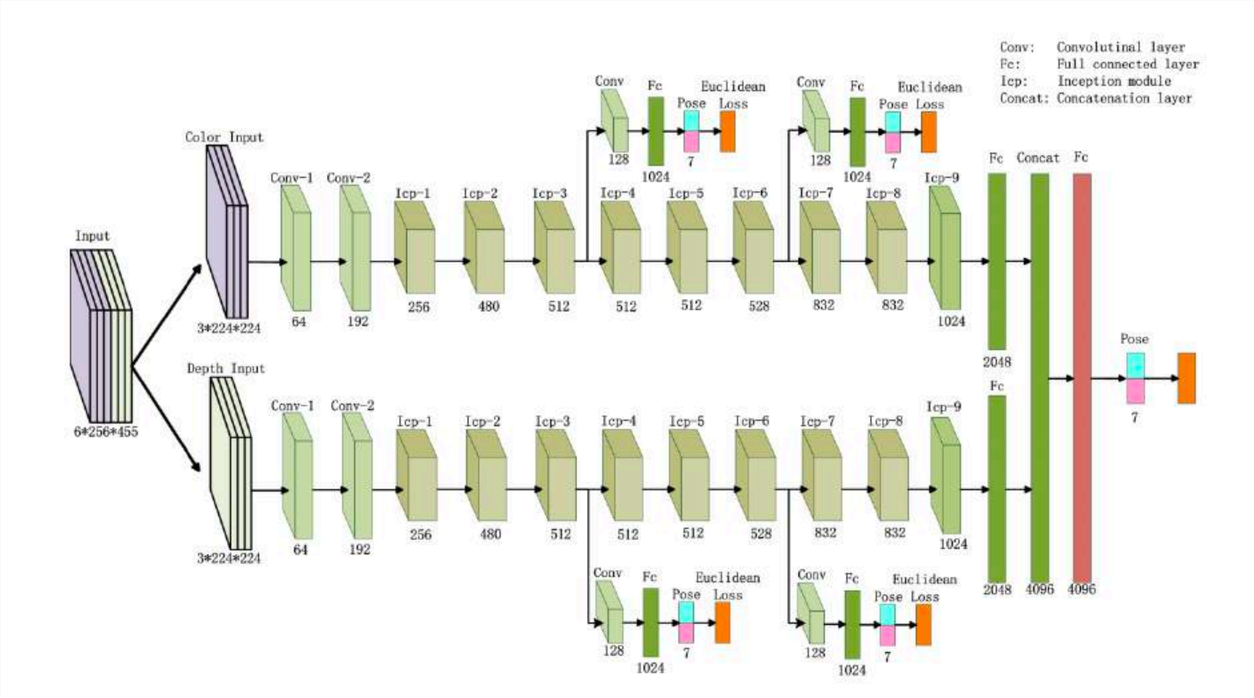
Video Localisation Improves Temporal Consistency

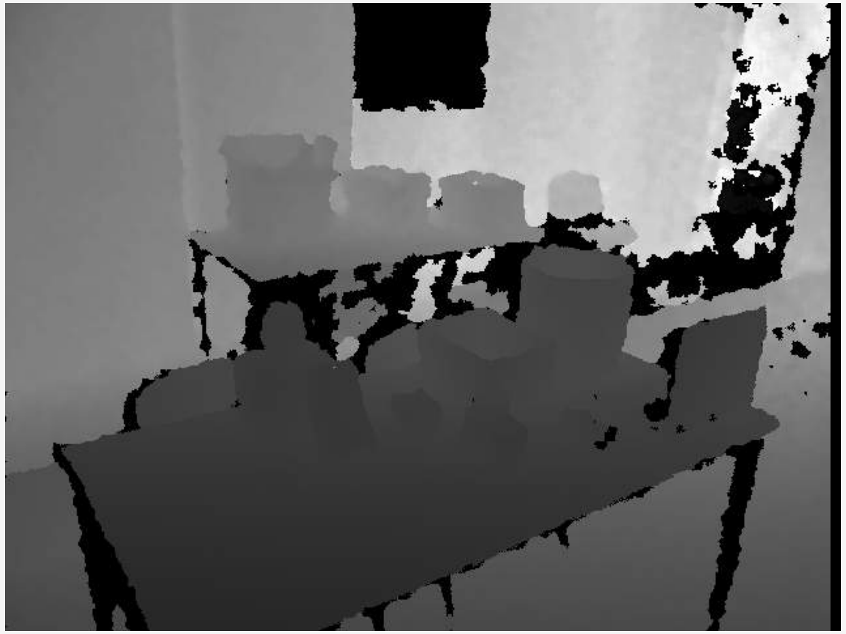
- Outperforms smoothing baseline
- Diminishing returns using very long sequences



Clark et al., ***VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization.*** IEEE CVPR 2017.

Camera Pose Regression with RGB-D Sensors





7 Scenes Data Results	Median Position	Median Orientation
PoseNet RGB	0.52m	12.8
Dual Stream PoseNet + Depth	0.35m	10.2

Ruihao Li et al. **Indoor relocalization in challenging environments with dual-stream convolutional neural networks.** *IEEE Transactions 2017*

Conclusions + Further Research Questions?

✓ Learning can produce **more efficient map** representations

✓ Convolutional nets are **more robust** to environmental challenges

? **Accuracy** still not as good as traditional feature methods

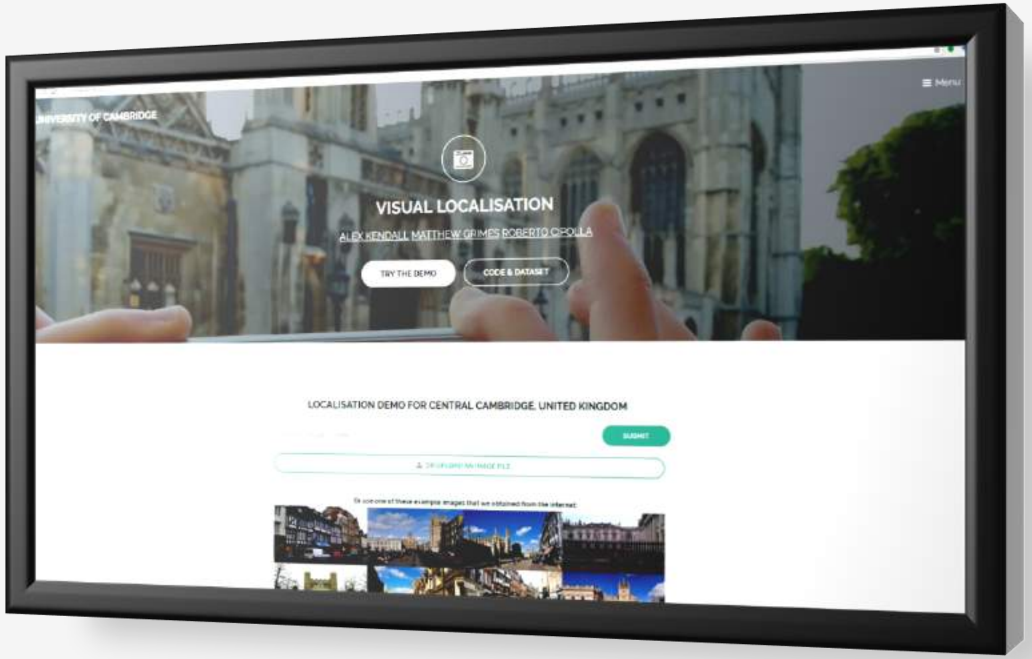
? How do we learn **city-scale localisation** map with deep learning?

? How can we **learn and update a map online**?

? How do we obtain **large city-scale datasets** with accurate pose labels?

Thank you! Questions?

- Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.
- Alex Kendall and Yarin Gal. **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?** arXiv preprint 1703.04977, 2017.
- Alex Kendall and Roberto Cipolla. **Geometric loss functions for camera pose regression with deep learning.** CVPR, 2017.
- Alex Kendall and Roberto Cipolla. **Modelling Uncertainty in Deep Learning for Camera Relocalization.** ICRA, 2016.



alexgkendall.com ← **Slides available soon!**

PoseNet Webdemo:
mi.eng.cam.ac.uk/projects/relocalisation/