CVSSP Centre for Vision, Speech and Signal Processing

Localisation via Deep Imagination: learn the features not the maps

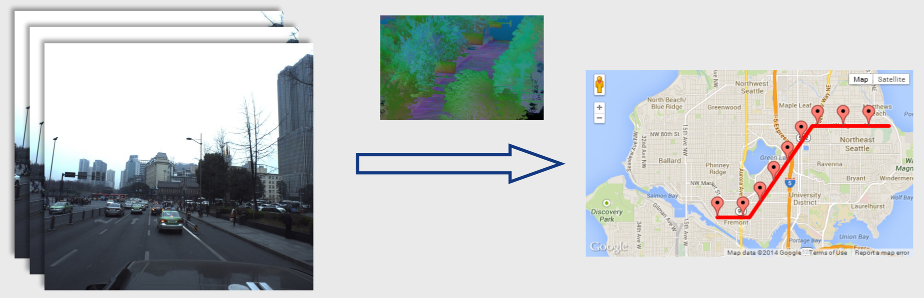
Jaime Spencer Martin (jaime.spencer@surrey.ac.uk) Oscar Mendez Maldonado (o.mendez@surrey.ac.uk) Richard Bowden (r.bowden@surrey.ac.uk) Simon Hadfield (s.hadfield.@surrey.ac.uk)

Centre for Vision Speech and Signal Processing, Faculty of Engineering and Physical Sciences, University Of Surrey, Guildford, Surrey, GU2 7XH

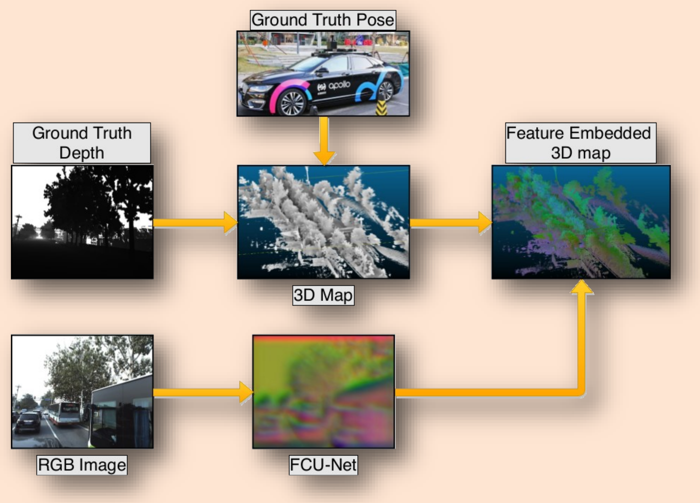
UNIVERSITY OF SURREY

1 - Abstract

- **Localisation** is fundamental to interacting with the world.
- Our work aims to perform **on-the-fly monocular localisation** using a prebuilt 3D map of the world.
- We propose a novel approach where an agent is able to model **new environments** after a **single visitation** by creating a **feature embedded 3D map**.
- The system uses the built map to **"imagine"** the scene's appearance from **novel, unseen poses**. These are used to estimate view likelihood and propagate the particles in a **Deep Imagination** localiser.

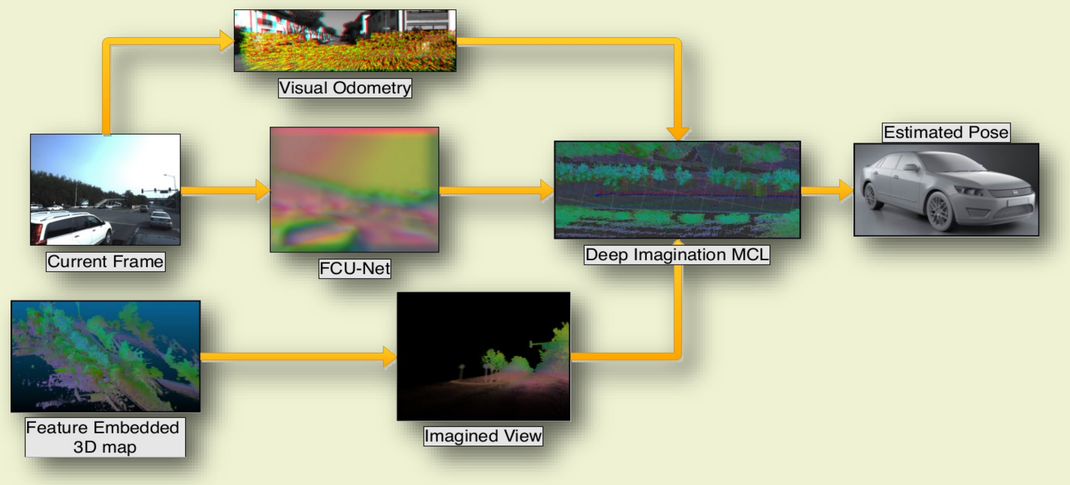


2 - Overview - Map building



- **Dense feature extraction** - Siamese **FCU-Net** pre-trained to produce **appearance invariant** descriptors.
- **Feature embedded 3D map** - **Backproject** FCU-Net **feature maps** into 3D space.

3 - Overview - Deployment



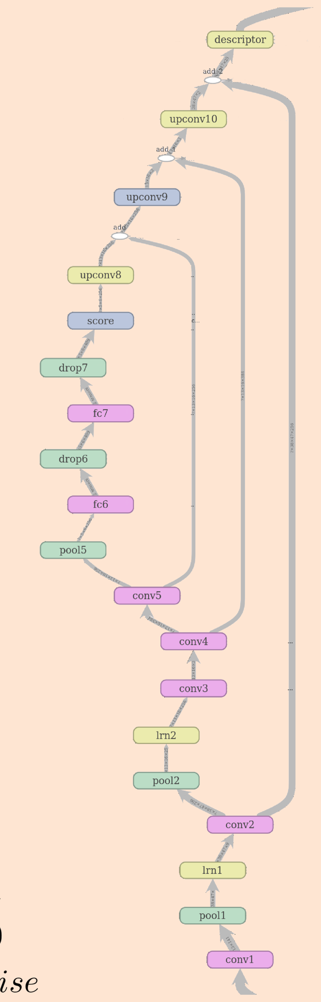
- **Deep Imagination localiser** - **VMCL**; **Project** feature embedded **map** to particle **pose hypothesis** to **"imagine"** the appearance and calculate view likelihood.
- **Visual odometry** - Particle filter **motion model**.

4 - FCU-Net

- **Dense** feature extractor
- **Fully convolutional** encoder-decoder
  - No size restrictions
- **Skip connections**
  - Combine low level **spatial** and high level **semantic features**

- Mapping $U : \mathbb{N}^3 \rightarrow \mathbb{R}^n$
- Training
  - **Siamese** network
  - **Pre-trained** on Kitti dataset
  - Novel **pixel-wise** contrastive loss:

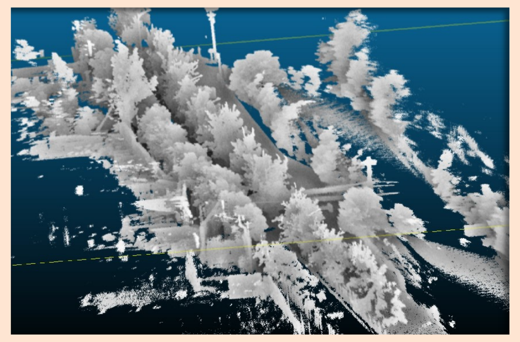
$$l(y, \boldsymbol{p}_1, \boldsymbol{p}_2) = \begin{cases} \frac{1}{2}(d)^2 & \text{if } y = 1 \\ \frac{1}{2}\{\max(0, m - d)\}^2 & \text{if } y = 0 \\ 0 & otherwise \end{cases}$$



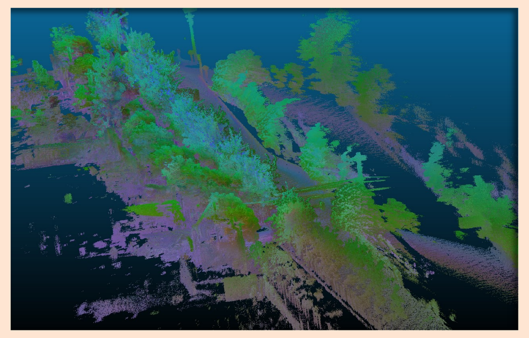
FCU-Net Architecture

5 - Feature embedded 3D map

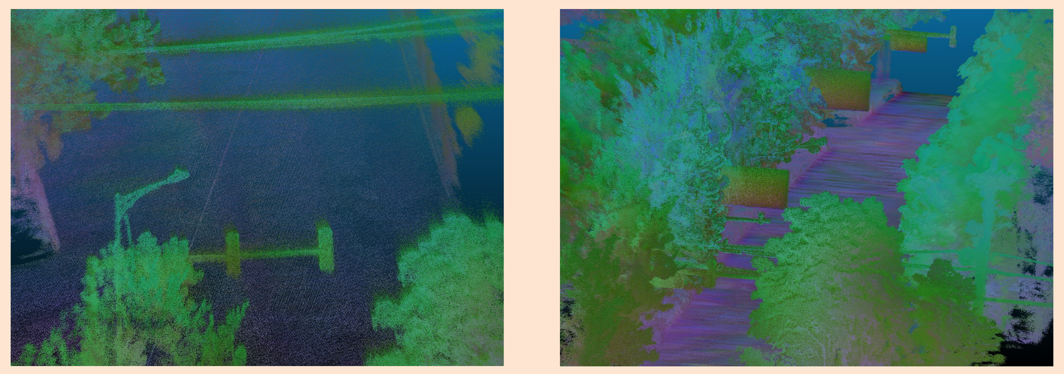
- **FCU-Net**
  - Dense feature extraction from each image
- **Backproject**
  - 3D location of each feature from ground truth pose and depth
- **Merge**
  - Final voxel descriptor is average of all candidates



Original Map



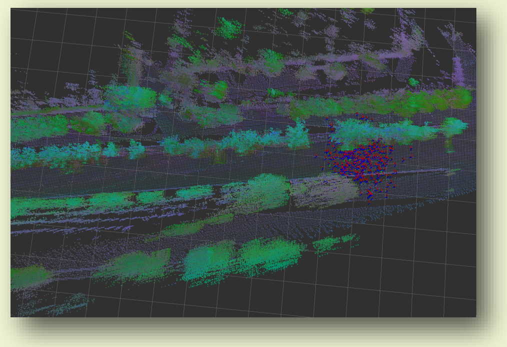
Feature embedded map



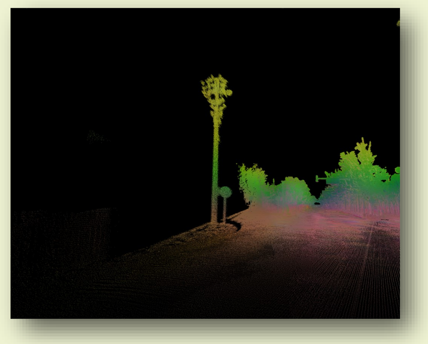
Embedded map details

6 - Deep Imagination localiser

- **Monte Carlo** localisation
  - Represent pose with 6 DOF
  - Each particle represents **one pose hypothesis**
- Particle **"imagined"** views
  - **Project** map onto each particle's **prior pose**
  - Take **depth** into account for **size and occlusion**

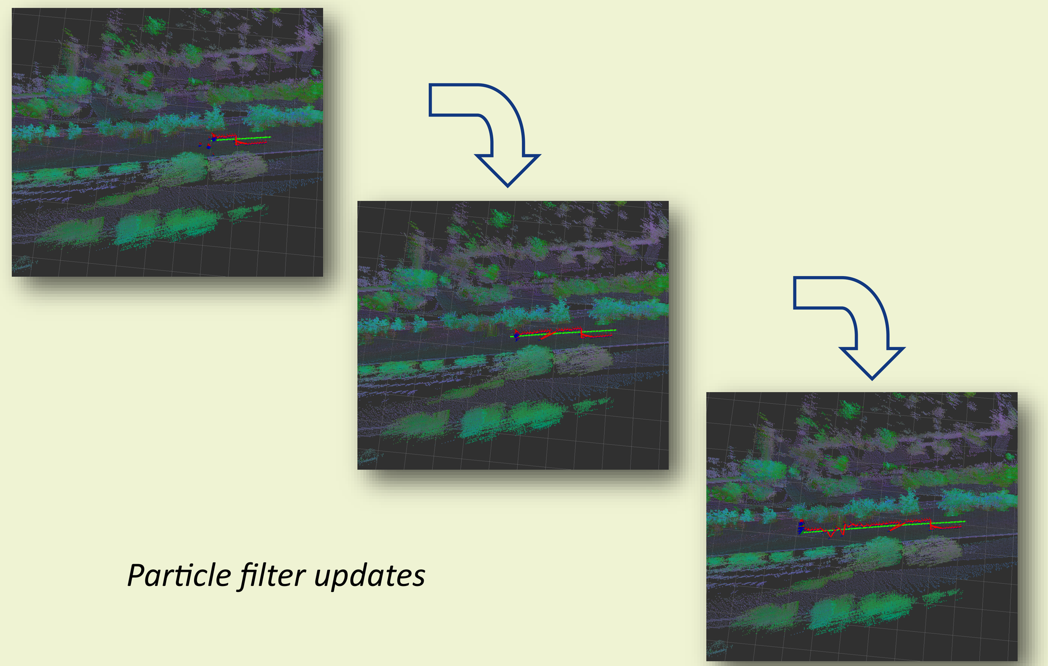


Particle filter initialization



Deep imagined view

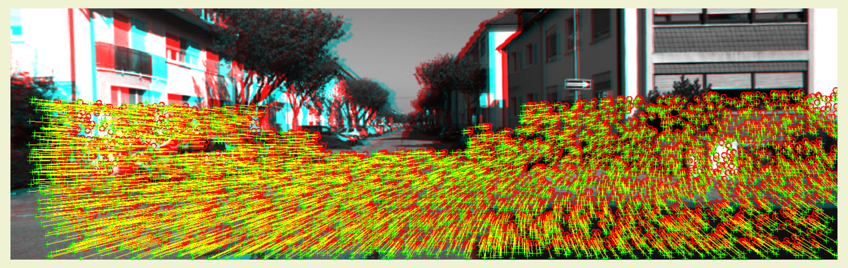
- Particle **likelihood** computation
  - **RMSE** between observation & imagined view
- Particle **weight update**
  - Combine likelihood and VO motion estimation
  - **Approximate MAP** encoding prior and likelihood



Particle filter updates

7 - Visual odometry

- **Essential matrix** estimation given I_t & I_{t-1}
  - **ORB** feature matching
- 4 possible **R**, **t** combinations
  - Combination with most points in front of camera



Sample dataset odometry

8 - Results

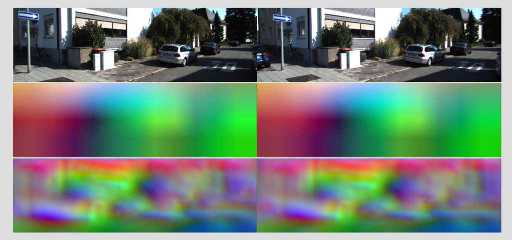
8.1 - Dense feature representation

- **Generic** features
  - Comparable results, **regardless of dataset trained on**

D	$Train$	$Test$	μ_{match}	$\mu_{non-match}$	$\mu_{overlap}$	D	$Train$	$Test$	μ_{match}	$\mu_{non-match}$	$\mu_{overlap}$
3			0.107	1.156	0.131	3			0.129	1.075	0.109
10	K	K	0.103	1.034	0.138	10	A	K	0.133	0.977	0.124
32			0.103	1.035	0.139	32			N/A	N/A	N/A
3			0.212	0.803	0.162	3			0.201	1.343	0.126
10	K	A	0.222	0.771	0.161	10	A	A	0.192	1.174	0.111
32			0.235	0.693	0.183	32			N/A	N/A	N/A

*Average distance between similar/dissimilar pairs. (**K**itti vs. **A**polloScape train/test)*

- Skip connections increases descriptor resolution

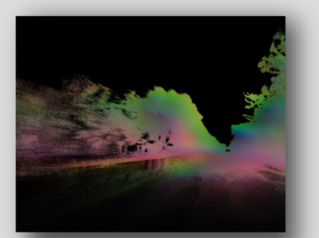


Top: Images - Mid: Base FCN - Bottom: FCU-N (skip)

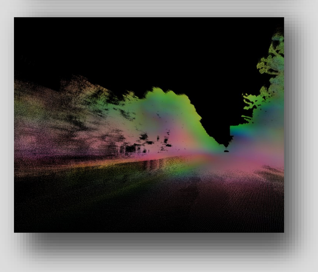
8.2 - Deep Imagination

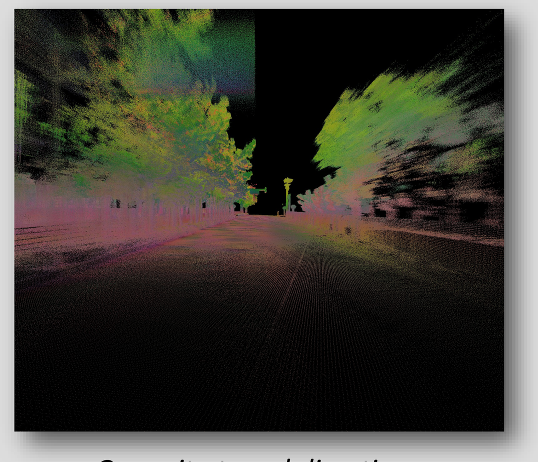
- **Novel imagined viewpoints**

Ground truth



Imagined





Opposite travel direction

- Sequence location prediction

Estimated path over ApolloScape validation sequence

9 - Conclusions

- **Deep Imagination** localiser generates feature representations for **unseen viewpoints.**
- Build a **representation of an environment** and imagine what it looks like from any new position **after a single visitation**.
- Generic features means **no additional training** is required in new environments
- Further improvements from a more sophisticated pose likelihood estimation (**PnP**) or additional motion models (**non-holonomic constraints**)

This work was funded by the EPSRC under grant agreements (EP/R512217/1) and (EP/R03298X/1) and Innovate UK Autonomous Valet Parking Project (Grant no 104273).

We would also like to thank NVIDIA Corporation for their Titan Xp GPU grant.

EPSRC Engineering and Physical Sciences Research Council