FocusTune: Tuning Visual Localization through Focus-guided Sampling

Son Tung Nguyen, Alejandro Fontan, Michael Milford, Tobias Fischer
QUT Centre for Robotics
Queensland University of Technology, Australia
✉ sontung.nguyen@hdr.qut.edu.au

1. The visual localization task

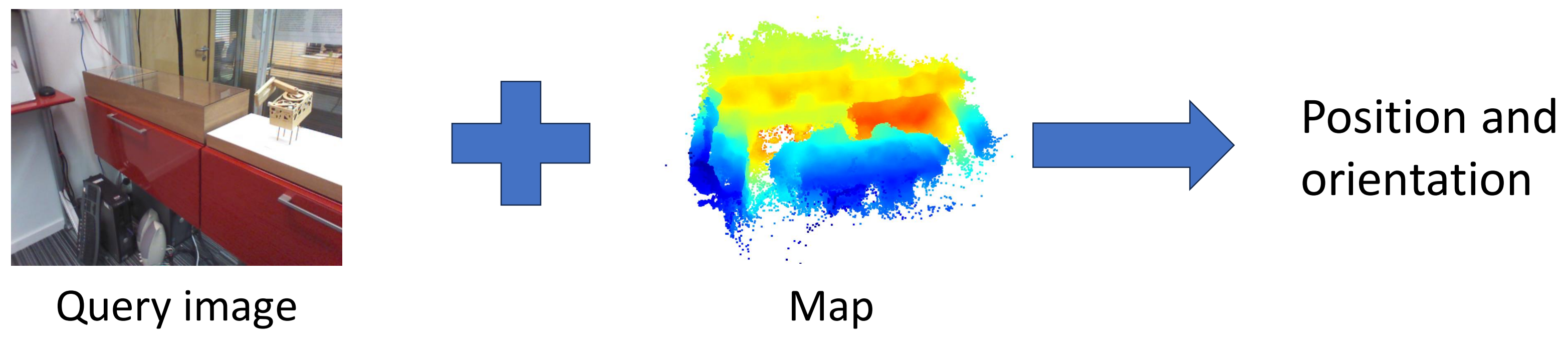


Figure 1: the overall task.

2. Solution basics

- 2D-3D correspondences are established between 2D features of the query image and 3D points of the map.
- These correspondences are then forwarded to a Perspective-N-Point solver to solve for the camera pose.

3. Related works

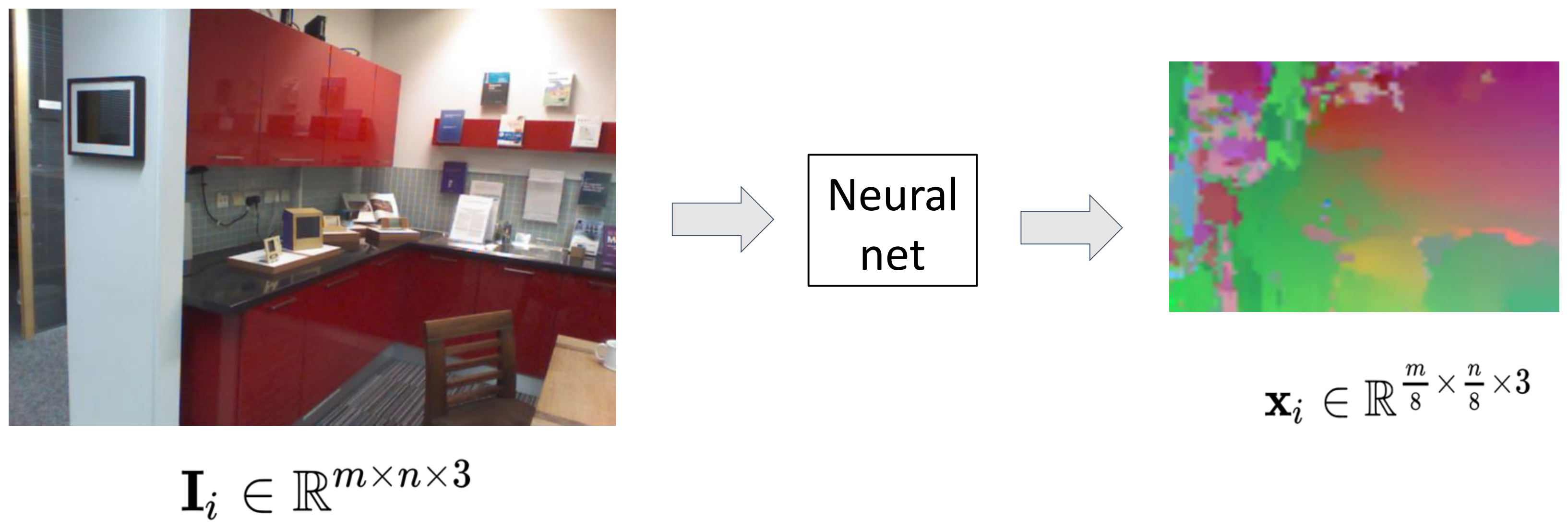

Neural net
$\mathbf{I}_i \in \mathbb{R}^{m\times n\times 3}$
$\mathbf{x}_i \in \mathbb{R}^{\frac{m}{8}\times\frac{n}{8}\times 3}$

Figure 2: Scene coordinate regression pipeline.

Scene coordinate regression methods (Fig. 2) [1, 2] are less accurate than structured methods [3], but:

- Are more memory efficient.
- Train fast (with ACE method).

4. Motivation

- Unique points are easier to triangulate *(top row Fig. 4 with the re-projection error of 0.74 pixels)*.
- Ambiguous points are harder to triangulate *(bot row Fig. 4 with the re-projection error of 72.6 pixels)*.
- Current methods uniformly sample all pixels, covering unique and ambiguous points in training *(Fig. 4 left figure)*.
- We propose to sample only the unique points and discard the ambiguous points *(Fig 4 right figure)*.

5. Methodology

- Re-project the point cloud onto the training image using the camera pose *(Fig. 5 top right image)*.
- Obtain the 2D re-projection as seed keypoints *(blue dots in Fig. 5 bottom right image)*.
- Sample training pixels uniformly within a circular region of 5 pixels surrounding the seed keypoints *(green region of the bottom-right image)* and prohibit sampling from outside these circles *(red region of the bottom-right image)*.
- Salient features *(the pink stars in the bottom left image of Fig. 5)* will be kept, while non-salient features *(the yellow crosses in the bottom left image of Fig. 5)* will be discarded.

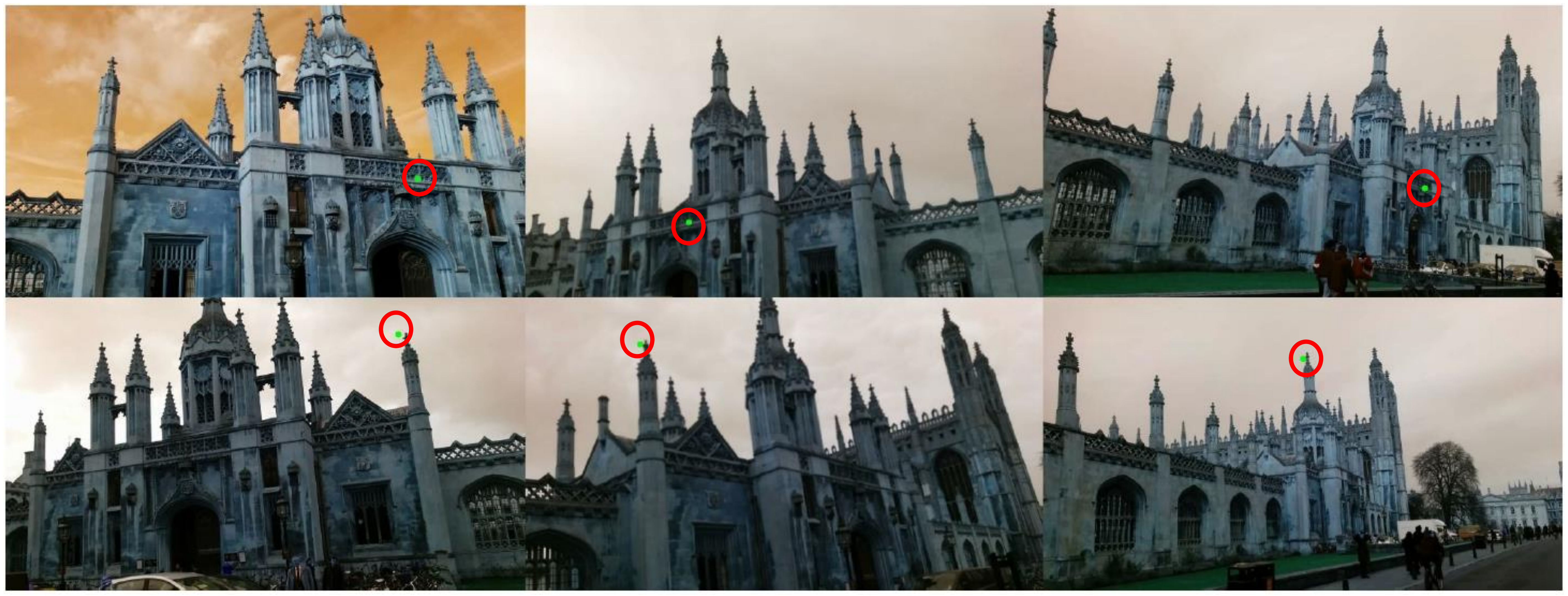
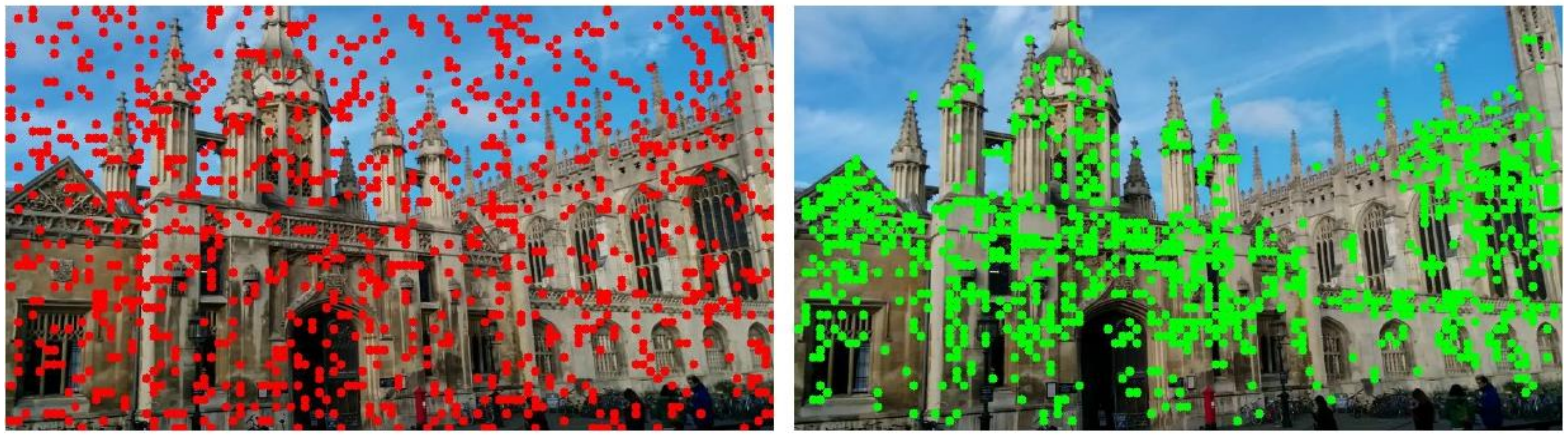


Figure 3: Examples of correct (top row with error of 0.74 px) and incorrect triangulations (bottom row with error of 72.6 px).



Random sampling (ACE)

Heuristic sampling (Ours)

Figure 4: Comparison of ACE [1] (left) where random sampling is used, and our heuristic FocusTune sampler (right).

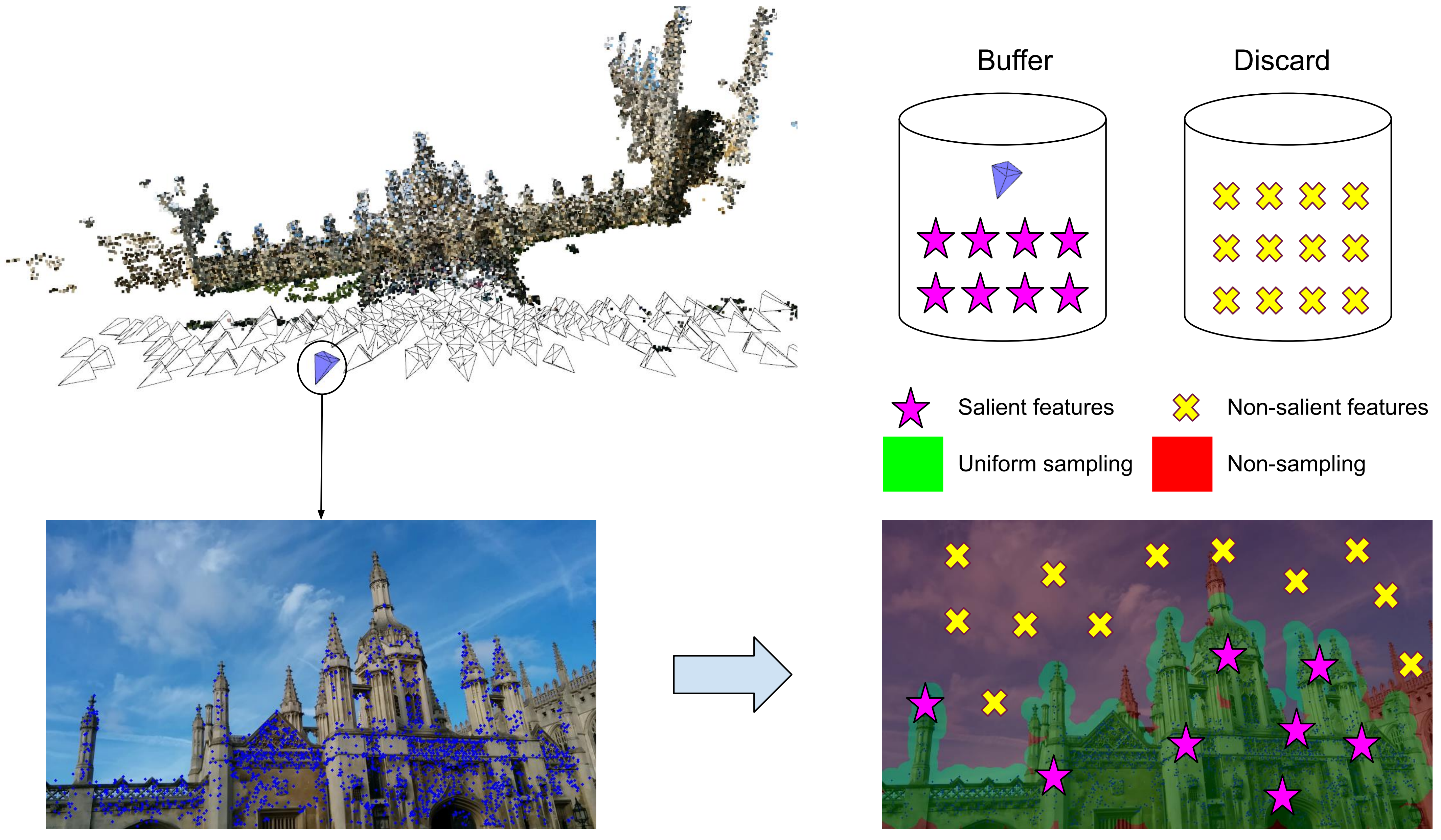


Figure 5: Re-projection of the map onto training images to obtain seed keypoints to specify uniform sampling regions to focus on salient features rather than non-salient features.

6. Results

7-scenes dataset
(percentage of test images under 5cm/5deg)

Active search	**98.5%**
ACE	97.2%
FocusTune (ours)	97.9%

Cambridge Landmarks dataset
(median error cm/deg)

Active search	**14/0.2**
ACE	25/0.4
ACE ensemble	17/0.3
FocusTune (ours)	19/0.3
FocusTune ensemble (ours)	15/0.3

References

1. Accelerated Coordinate Encoding: Learning to Relocalize in Minutes using RGB and Poses, Brachmann et al., CVPR 2023.
2. Visual Camera Re-Localization From RGB and RGB-D Images Using DSAC, Brachmann et al., TPAMI 2021.
3. Improving Image-Based Localization by Active Correspondence Search, Sattler et al., ECCV 2012.