Visible Structure Retrieval for Lightweight Image-Based Relocalisation

Fereidoon Zangeneh[1,2] (fereidoon.zangeneh@univrses.com), Leonard Bruns[1], Amit Dekel[2], Alessandro Pieropan[2], Patric Jensfelt[1]

[1]KTH Royal Institute of Technology, [2]Univrses AB

The 36th British Machine Vision Conference

Introduction

Keypoints/points described by local features

- ✓ Locally informative
- ✗ Globally ambiguous

This makes feature matching difficult in large scenes.

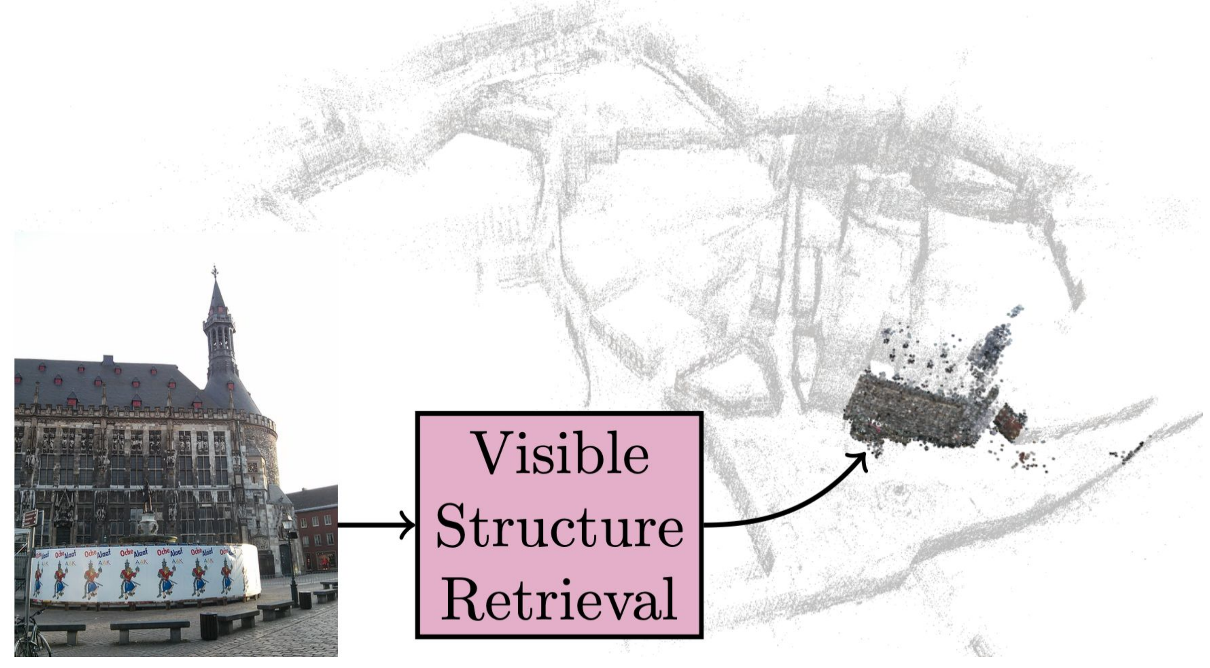
Remedies:

- Correspondences search heuristics [1]
  → results in complex pipelines
- Image retrieval in a hierarchical pipeline [2, 3, 4]
  → explicitly storing past observations grows with the number of observations

In this work, we propose implicitly modelling the scene in a lightweight neural network (taking a couple of megabytes), such that its forward pass efficiently reduces the search space for effective feature matching in large scenes.

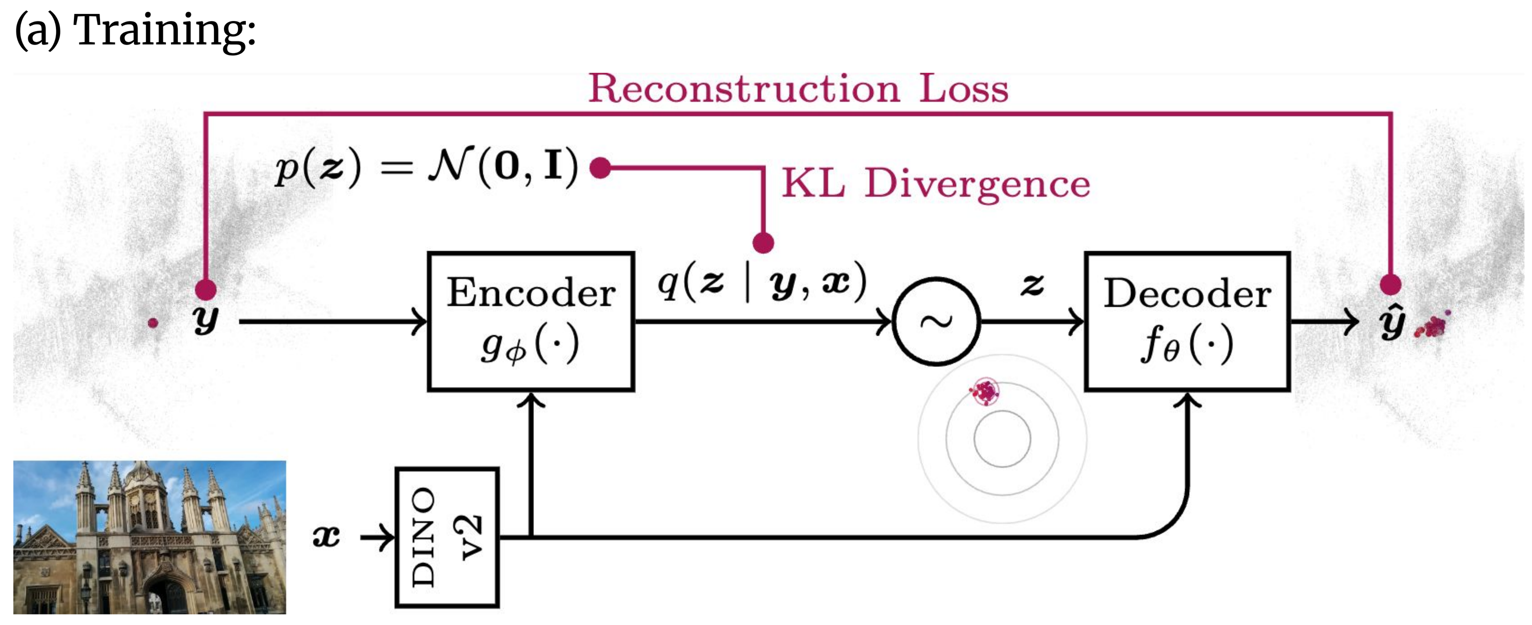
Visible Structure Retrieval

Our proposed method retrieves the subset of SfM points that are visible per image observation. This serves to reduce the search space for establishing 2D-3D correspondences in a structure-based localisation pipeline.

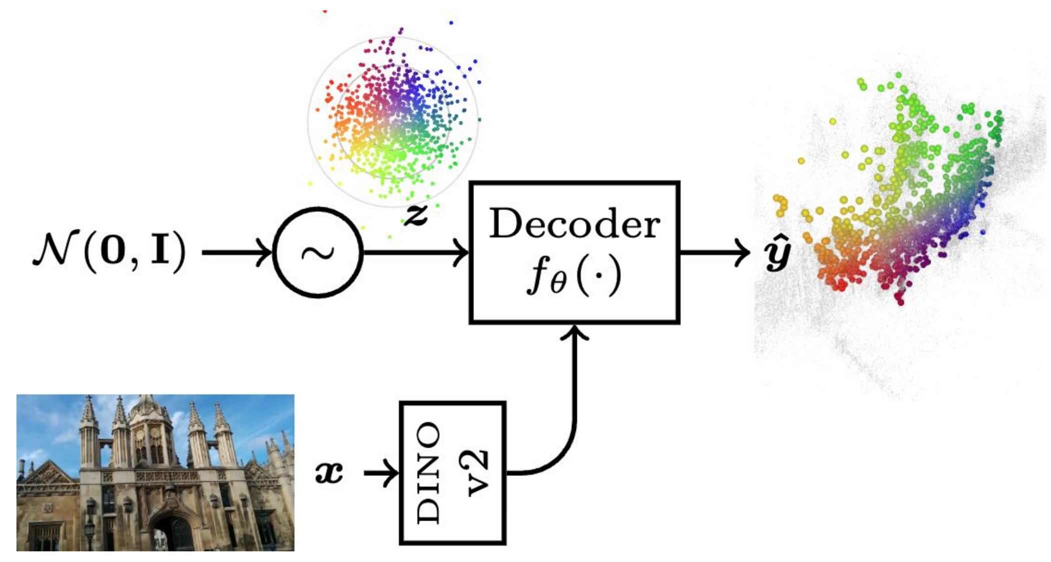


Generative Modelling of Structure Points

(a) Training:



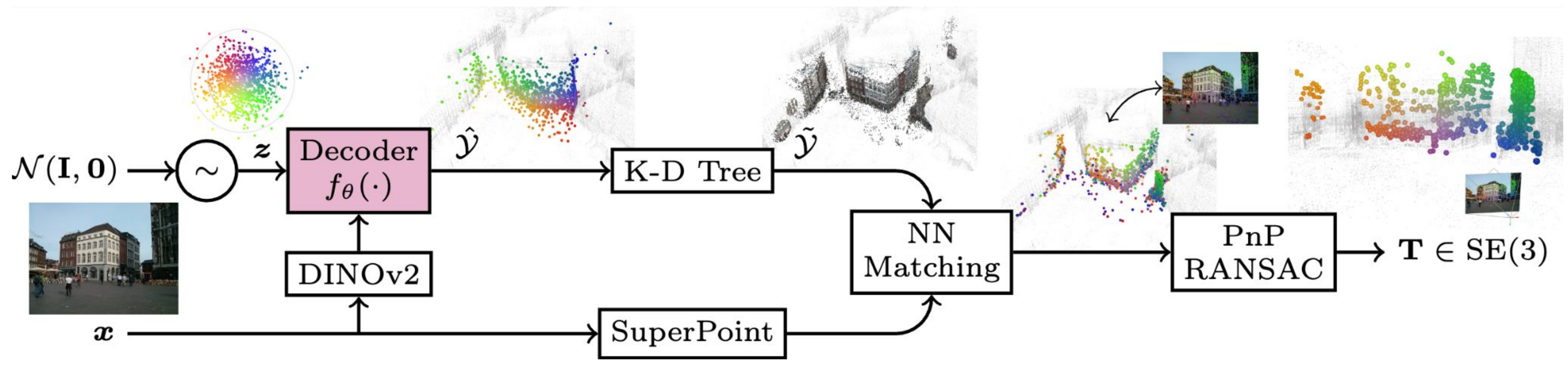
(b) Inference:



(a) Training: Train a VAE reconstructing individual SfM points, conditioned on reference images they're observed from.

(b) Inference: Decoding samples from the prior distribution generates 3D points spanning regions of the map that are visible in the conditioning image.

Lightweight Image-Based Relocalisation



(1) The visible structure in a query image is regressed from noise samples through a forward pass of our network.

(2) The SfM points are retrieved by a k-d tree radius search around the regressed points.

(3) The keypoints on the query image are matched with the retrieved 3D points through nearest neighbour search in descriptor space.

(4) Camera pose is estimated through PnP within a RANSAC framework.

Examples

Our approach effectively retrieves the set of 3D points observed in a query image. Using a pretrained feature extractor, such as DINOv2, to extract image-level embeddings as the conditioning vector enables achieving invariance to lighting conditions.

Results

Visible structure retrieval enables structure-based relocalisation in large-scale scenes, with an accuracy comparable to the state of the art, while requiring a smaller memory footprint. It replaces the explicit image retrieval database in a hierarchical pipeline with an implicit model that efficiently and effectively reduces the search space for feature matching.

Method	Aachen Day			Aachen Night			Storage for Retrieval	
	0.25m / 2°	0.5m / 5°	5m / 10°	0.25m / 2°	0.5m / 5°	5m / 10°	Size / MB	Complexity
HLoc (SP+SG)	89.8	96.1	99.4	77.0	90.6	100.0	56	$O(n)$
Active Search	85.3	92.2	97.9	39.8	49.0	64.3	–	–
NetVLAD (SP+NN)	85.2	92.7	97.6	67.0	83.2	92.7	56	$O(n)$
AnyLoc (SP+NN)	85.2	92.8	98.1	67.5	82.7	96.3	658	$O(n)$
MS-Transformer	0.0	0.0	0.0	0.0	0.0	0.0	72	$O(1)$
GLACE	8.6	20.8	64.0	1.0	1.0	17.3	27	$O(1)$
Ours: ViStR (SP+NN)	82.3	90.7	95.6	64.4	77.5	89.0	7	$O(1)$

References

[1] Sattler, T. *et al.* Efficient & effective prioritized matching for large-scale image-based localization. TPAMI 2016.

[2] Arandjelovic, R. *et al.* NetVLAD: CNN architecture for weakly supervised place recognition. CVPR 2016.

[3] Keetha, N. *et al.* AnyLoc: Towards universal visual place recognition. RAL 2023.

[4] Sarlin, P. E. *et al.* From coarse to fine: Robust hierarchical localization at large scale. CVPR 2019.

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM