CVPR VIRTUAL JUNE 19-25

BACK TO THE FEATURE

Learning Robust Camera Localization from Pixels to Pose

Paul-Edouard Sarlin[1*] Ajaykumar Unagar[1*] Måns Larsson[2] Hugo Germain[3]
Carl Toft[2] Viktor Larsson[1] Marc Pollefeys[1,4] Vincent Lepetit[3]
Lars Hammarstrand[2] Fredrik Kahl[2] Torsten Sattler[2,5]

ETH zürich
*Equal contributions

CHALMERS UNIVERSITY OF TECHNOLOGY

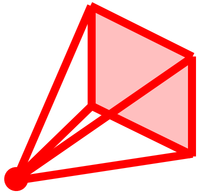
École des Ponts ParisTech

Microsoft

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE
CIIRC

6-DoF camera pose estimation

query

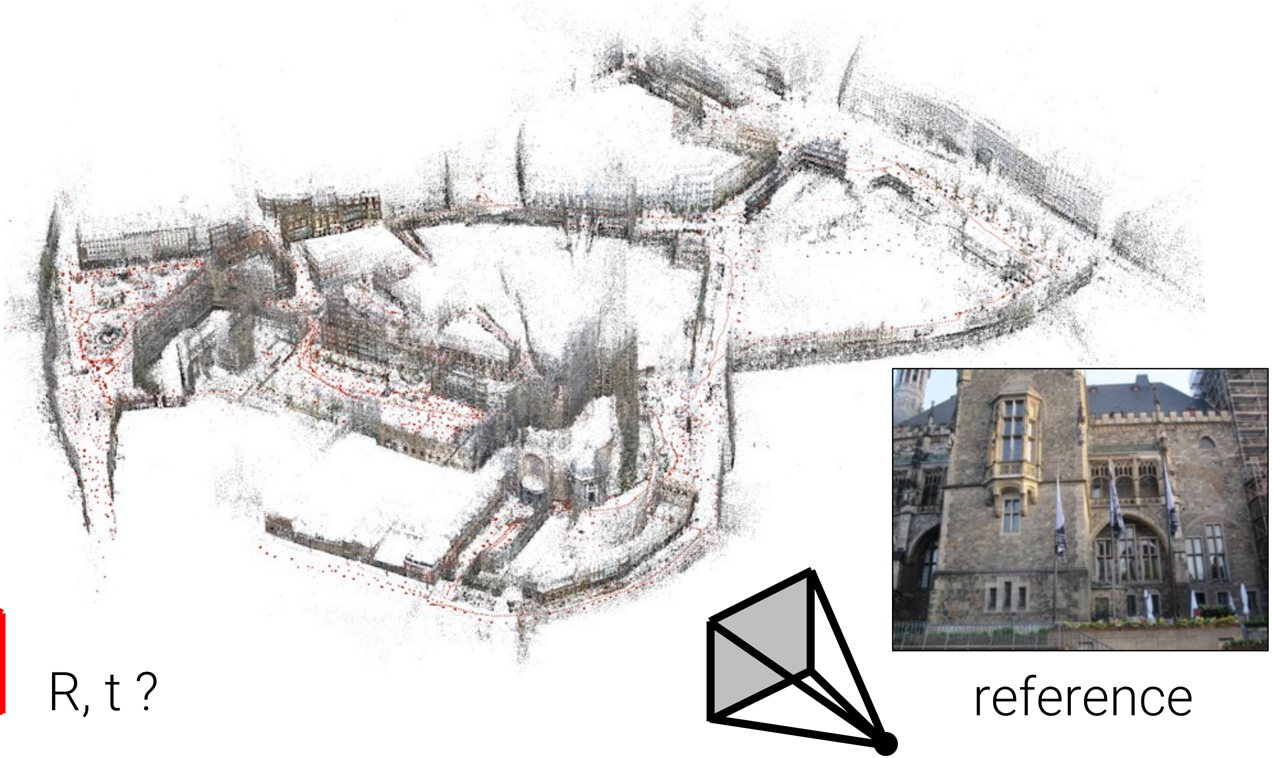


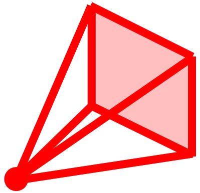
R, t ?

We tackle the task of visual localization, which estimates the rotation and the translation of a camera in a 3D environment given a single query image.

6-DoF camera pose estimation

query





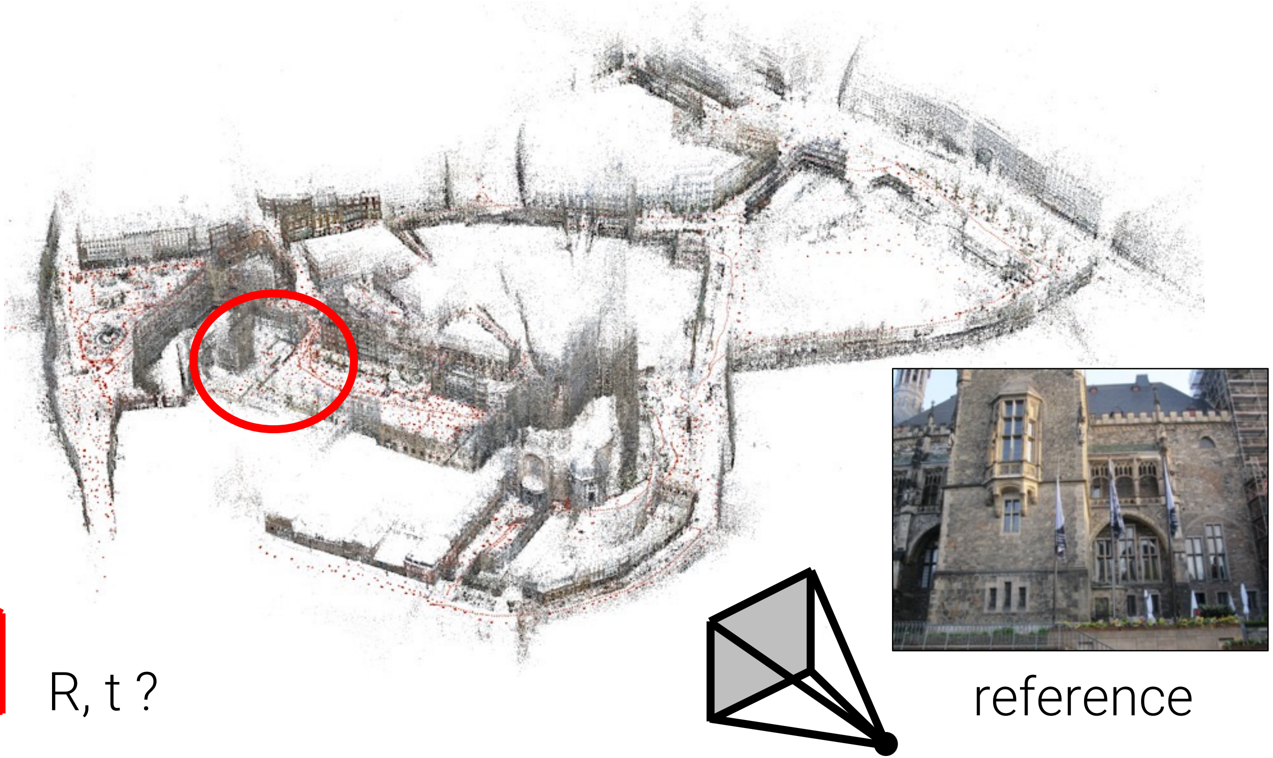
R, t ?

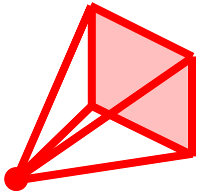
reference

We assume to have a map of the environment, composed of reference images with poses, and a 3D model (e.g. sparse pointcloud built with Structure-from-Motion).

6-DoF camera pose estimation

query



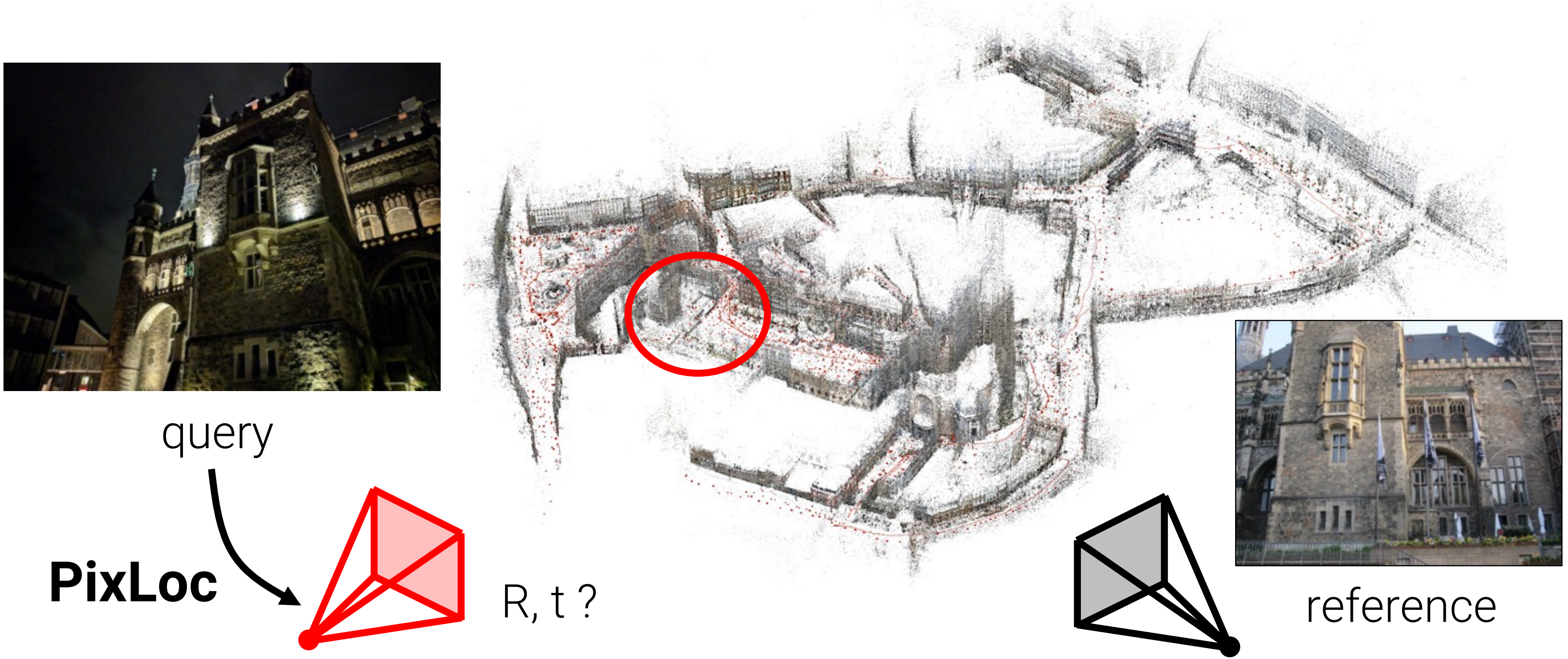


R, t ?

reference

We often also have a coarse prior on the pose, obtained with image retrieval or even GPS.

6-DoF camera pose estimation



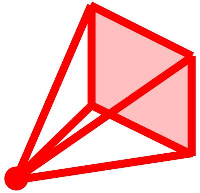
The contribution of this work is PixLoc, a learning algorithm that estimates the pose of a given image.

Currently: generalization *or* end-to-end

local feature matching

detection → description → matching → solver

query



pose

- ✅ Scene agnostic
- ✅ Good generalization
- ✅ Interpretable
- ❌ Complex pipeline

Current approaches to this problem belong to two categories. The classical pipeline detects local features, describes and matches them, and finally solves for the pose. Multiple of these blocks can be learned, but training end-to-end is difficult.

Currently: generalization *or* end-to-end

query

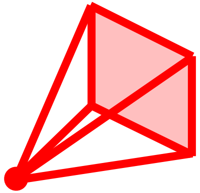
local feature matching

detection | description | matching | solver

Feature extraction | 3D point regression | solver

pose regression

end-to-end localization



pose

- ❌ Trained for each scene
- ❌ Poor generalization
- ❌ Blackbox
- ✅ Trained end-to-end

Instead, recent approaches like DSAC rely on a single convolutional neural network (CNN) to regress geometric quantities like 3D points. The CNN recognizes specific scene features and predicts their 3D coordinates or the corresponding viewpoint.

Currently: generalization *or* end-to-end

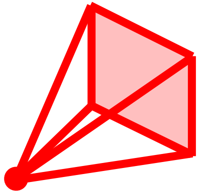
local feature matching

detection | description | matching | solver

query

Feature extraction | 3D point regression | solver

pose regression



pose

end-to-end localization

- ❌ Trained for each scene
- ❌ Poor generalization
- ❌ Blackbox
- ✅ Trained end-to-end

The weights of the CNN therefore encode the 3D structure of the reference views, and often cannot generalize to new scenes.
Some works instead regress poses relative to reference images, which in theory is not bound to a specific scene, but in practice still fails to generalize.

Currently: generalization *or* end-to-end

local feature matching

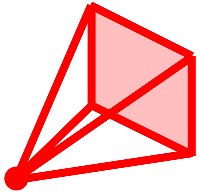
detection | description | matching | solver

query

Feature extraction | ~~3D point regression~~ | solver

~~pose regression~~

end-to-end localization



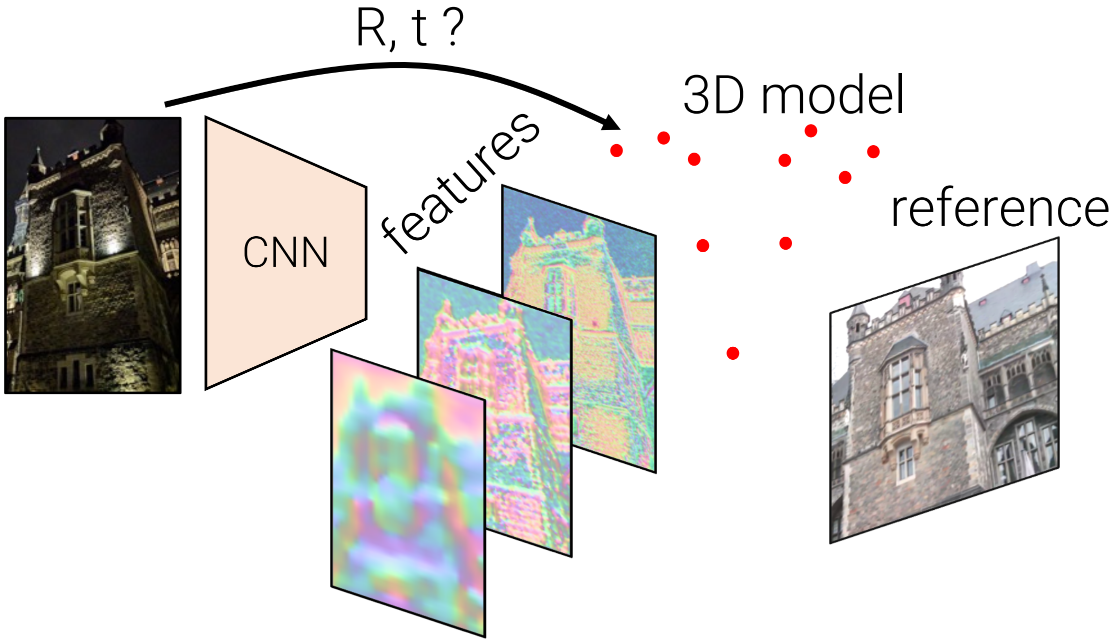
pose

use classical 3D geometry!

In this paper, we argue that deep neural networks do not need to learn 3D geometry. Instead, deep nets should go Back to the Feature: they only need to learn good features, and the regression should be performed with classical geometry.

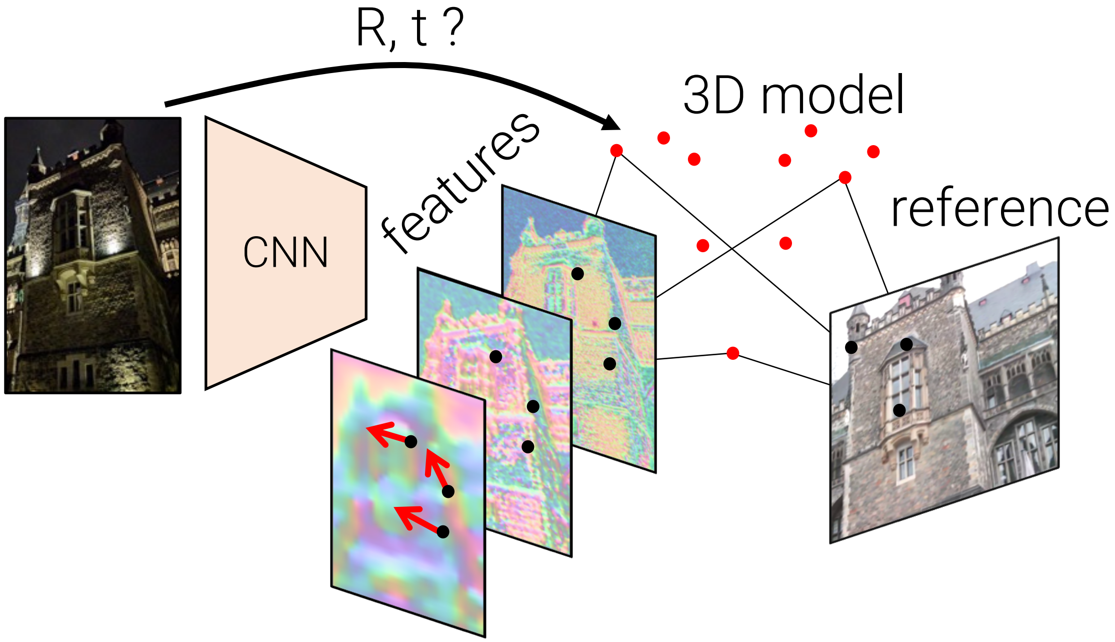
BACK TO THE FEATURE

Our approach: PixLoc



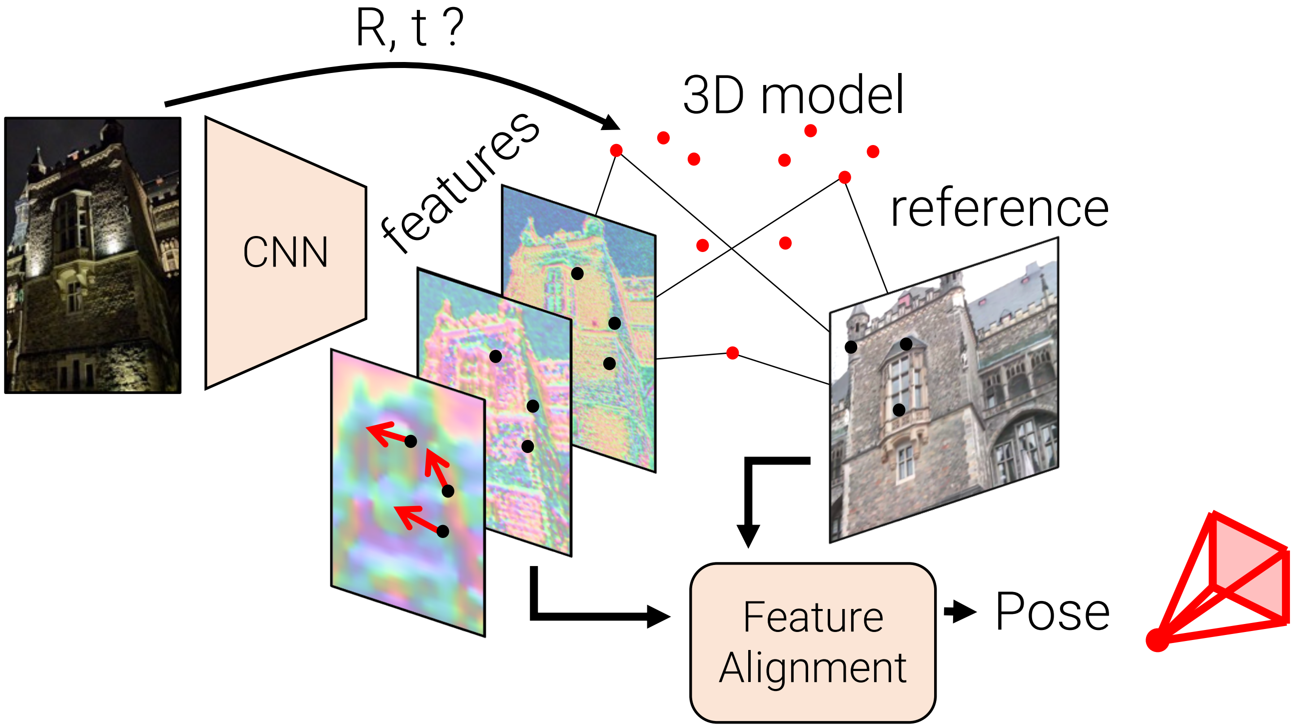
Let's have a closer look at PixLoc.
A CNN first predicts dense features for the query and for a corresponding reference image.

Our approach: PixLoc



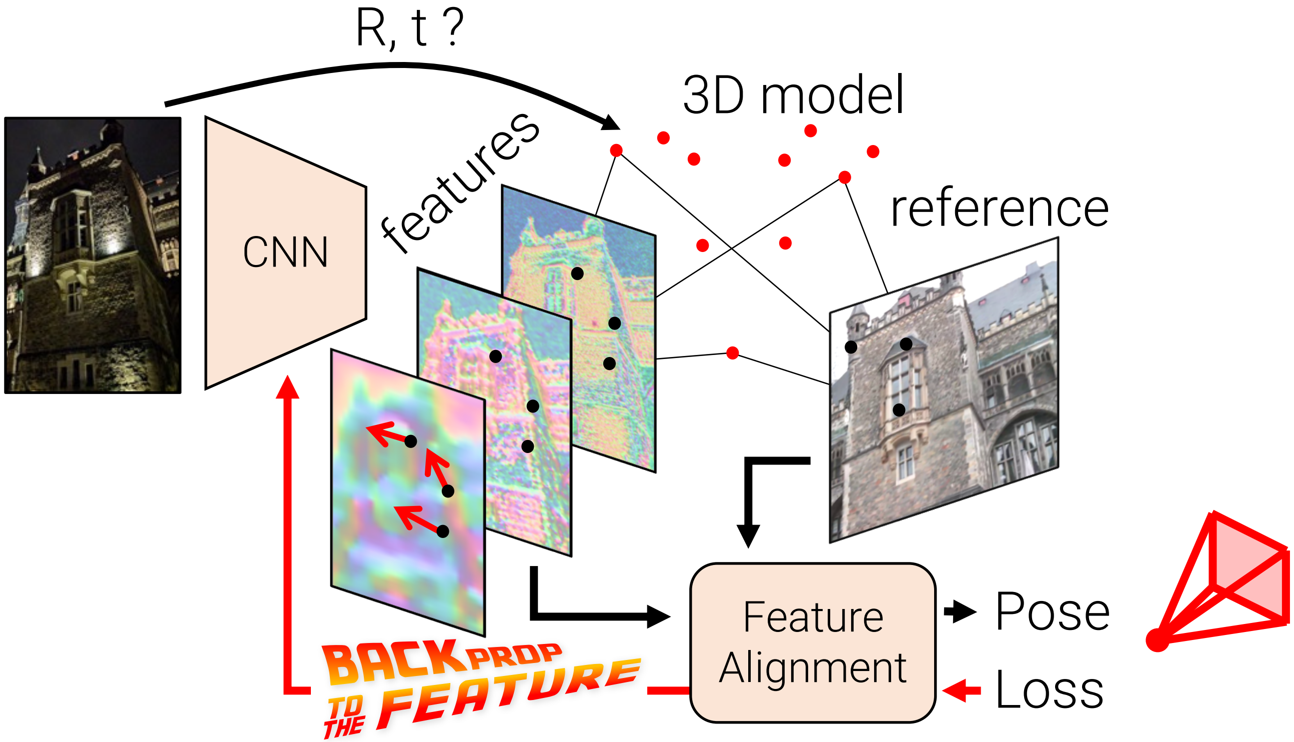
Given local 3D points and a coarse initial pose,
we can compute the error between query and reference features.

Our approach: PixLoc



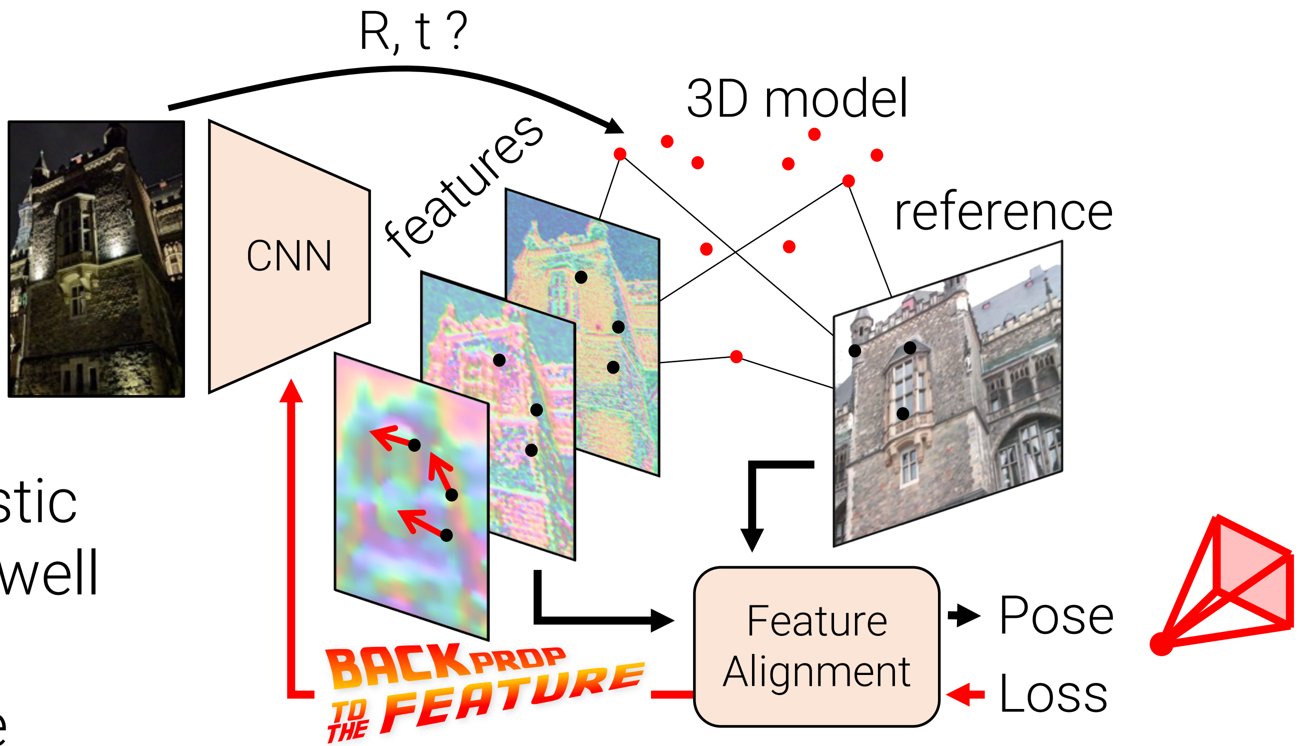
A geometric optimization then refines the pose by aligning the features.

Our approach: PixLoc



The optimization is differentiable so that PixLoc is trained end-to-end by backpropagating to the features.

Our approach: PixLoc

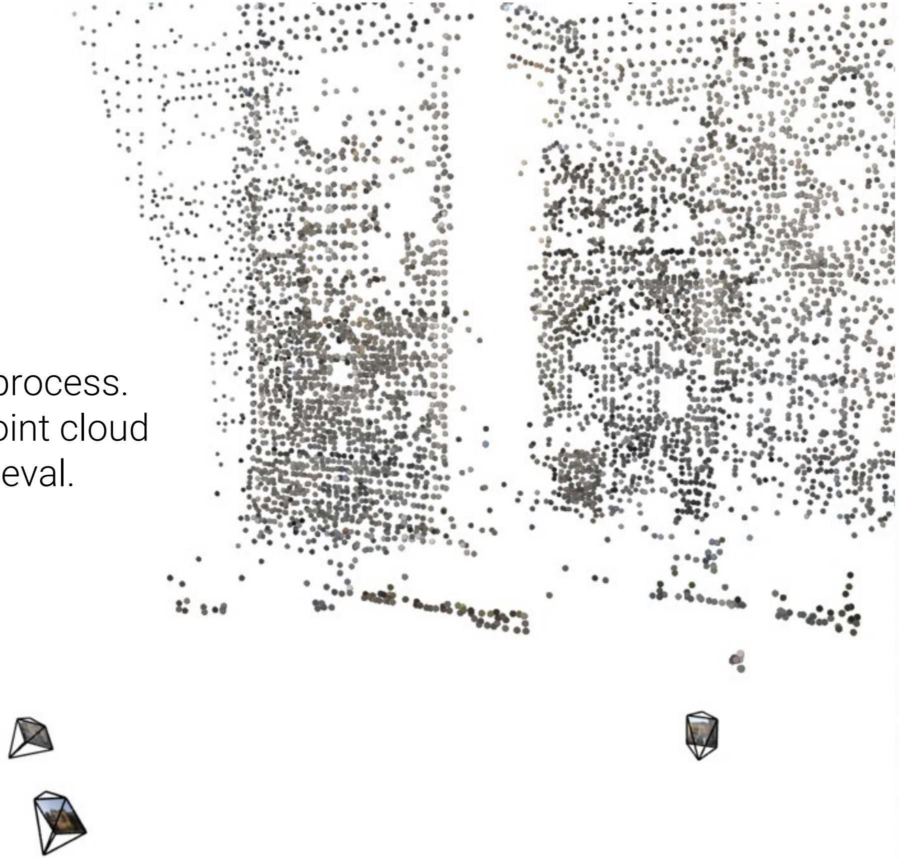


- ✅ Scene agnostic
- ✅ Generalizes well
- ✅ Accurate
- ✅ Interpretable
- ✅ Trained end-to-end

By taking the 3D information out the network, the features are generic.

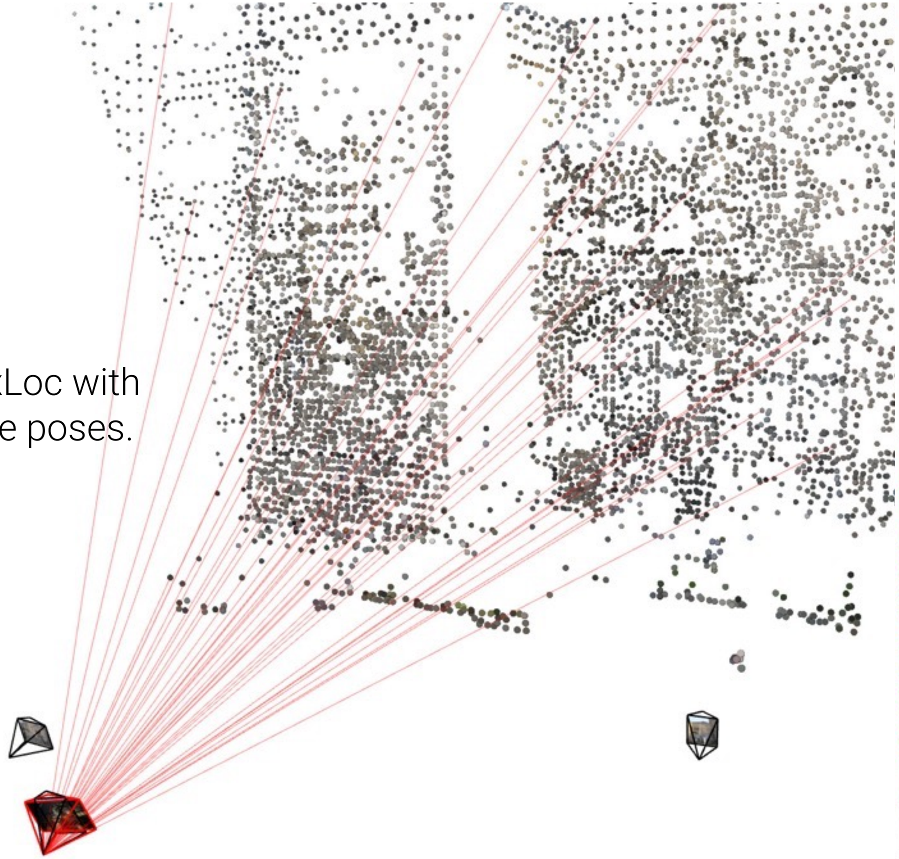
PixLoc: localization by image alignment

Let’s visualize the process.
We isolate a local point cloud
with image retrieval.



PixLoc: localization by image alignment

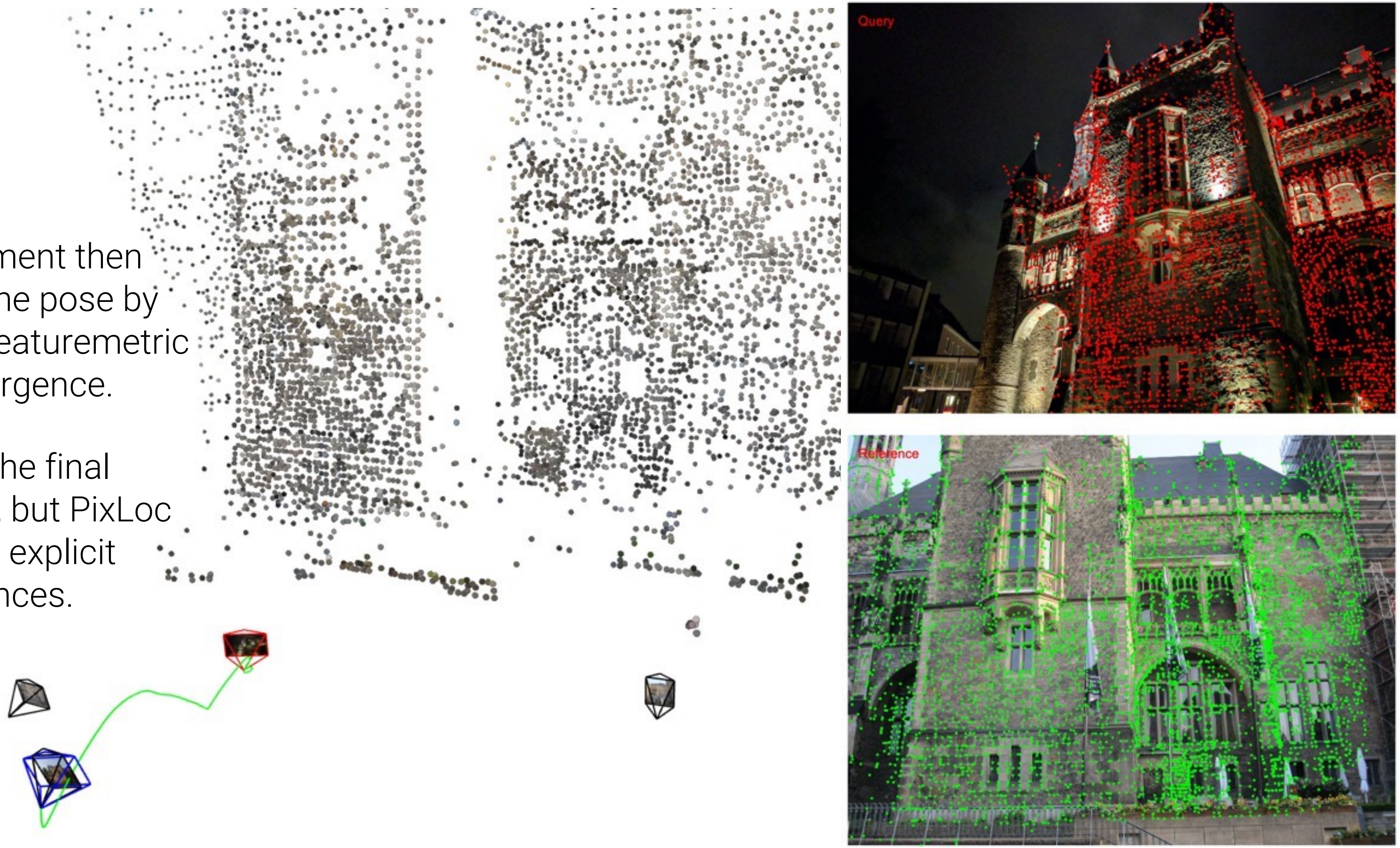
And we initialize PixLoc with one of the reference poses.



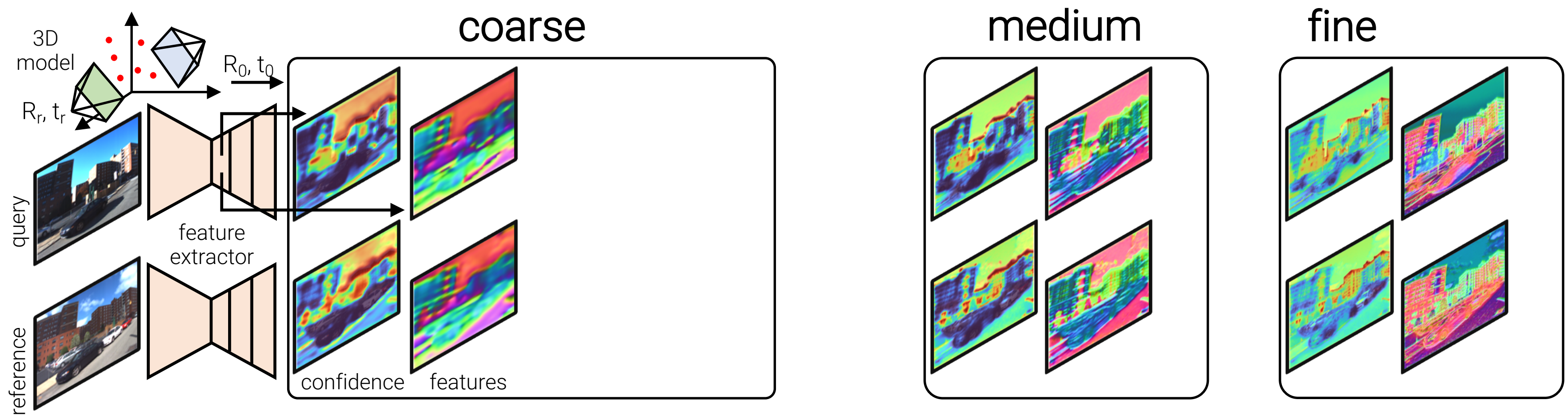
PixLoc: localization by image alignment

The feature alignment then iteratively refines the pose by minimizing a direct featuremetric cost until convergence.

Here we show the final reprojections in red, but PixLoc does not rely on explicit correspondences.

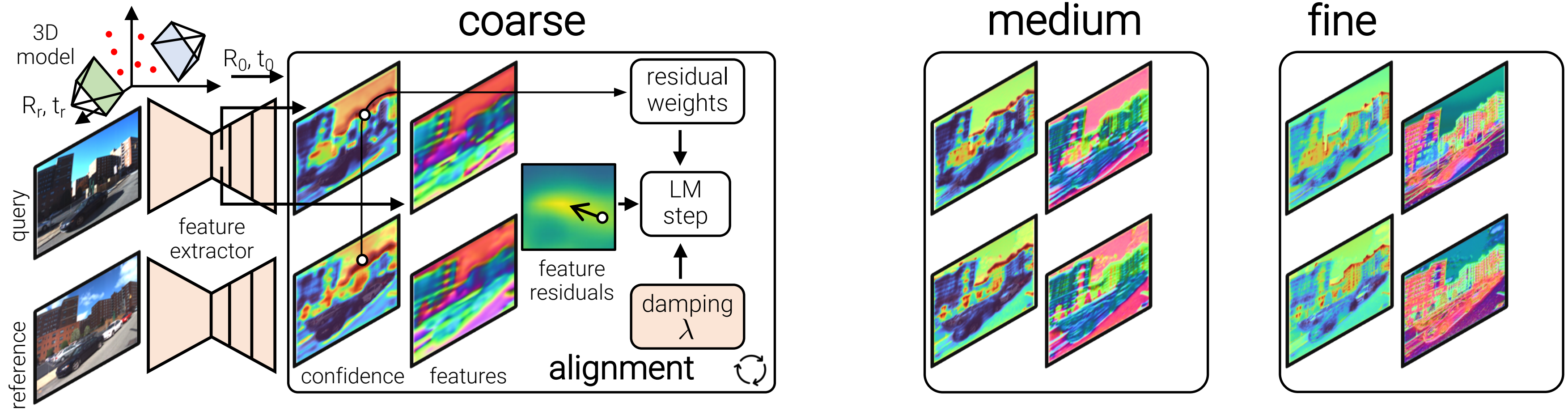


Multi-level optimization



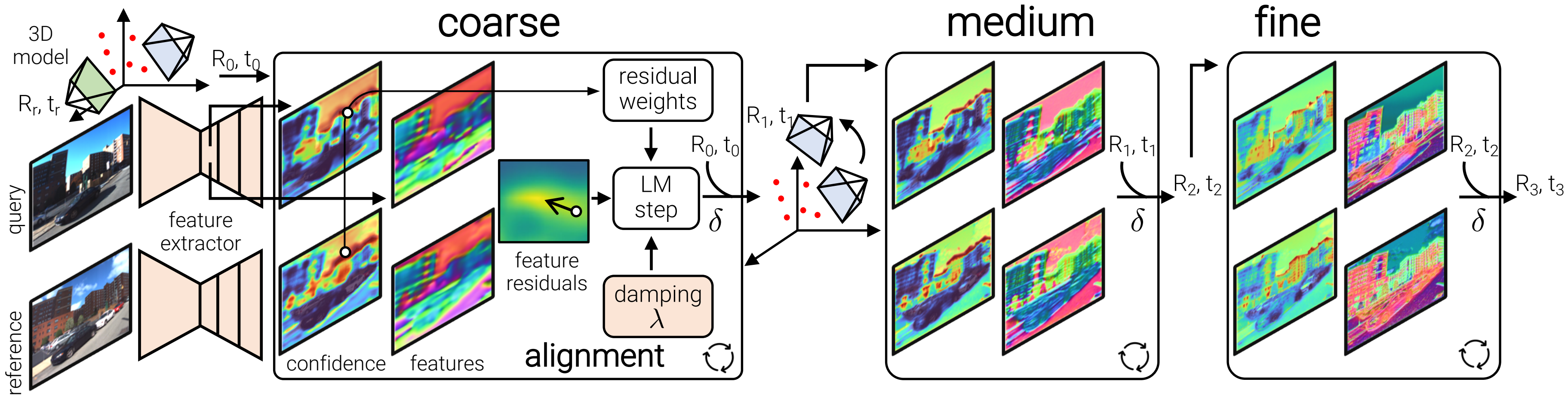
For each image, PixLoc first extracts dense features and corresponding confidence maps at multiple levels, from coarse to fine.

Multi-level optimization



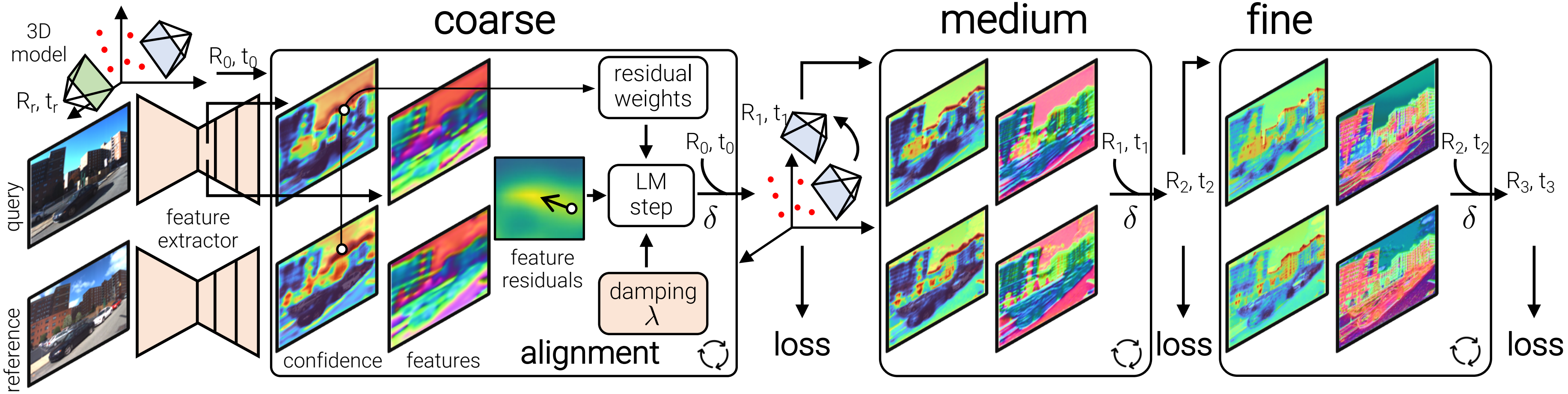
For each 3D point, the features define a cost that is weighted by the confidence and minimized using gradient-based optimization.
PixLoc also encodes a regularization λ that reflects the prior on the camera motion.

Multi-level optimization



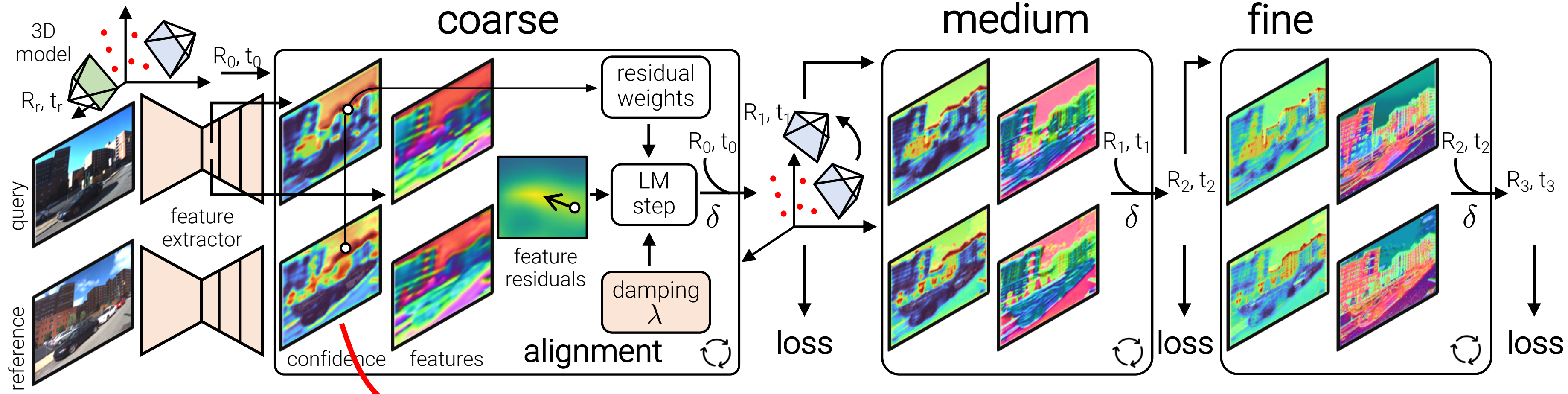
We obtain an updated pose, which initializes the optimization at the next level, and so on.

Multi-level optimization

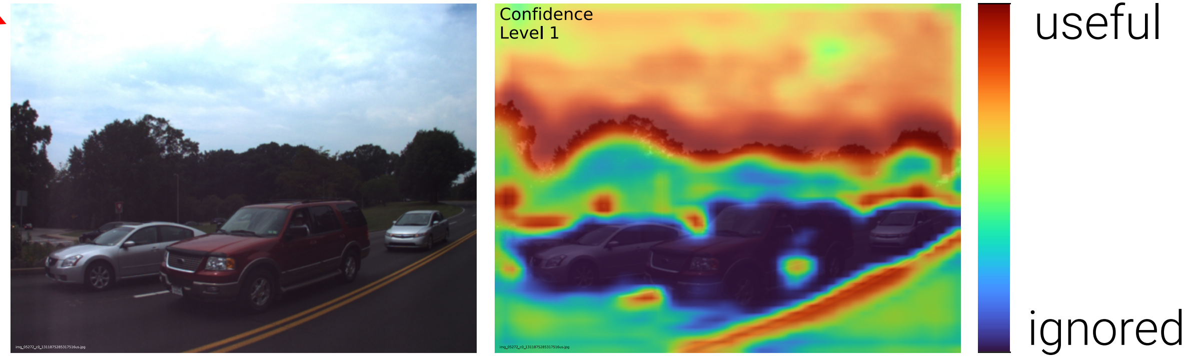


PixLoc is trained by supervising only the final poses, and thus does not require ground truth 3D geometry.

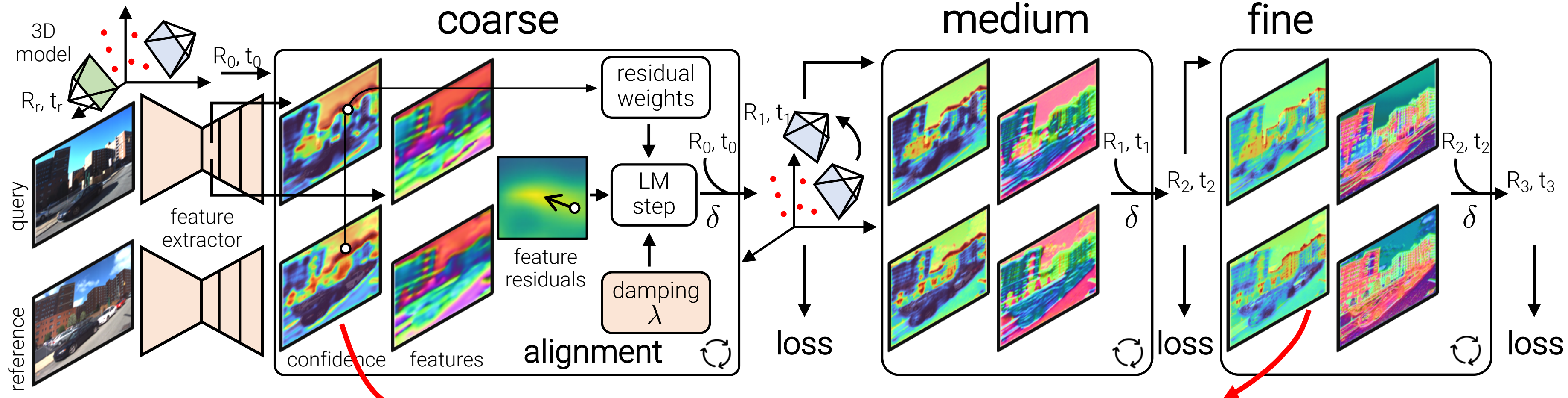
Multi-level optimization



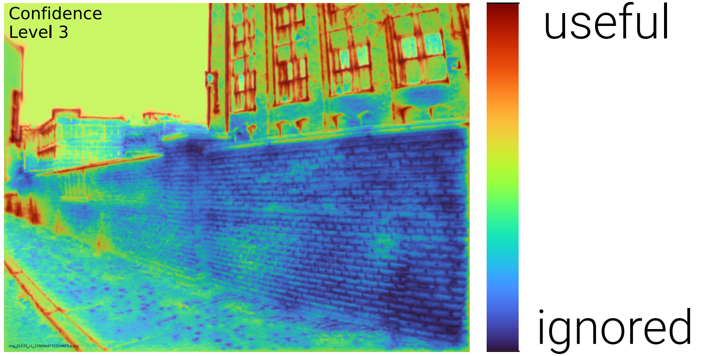
The confidence learns to ignore dynamic objects like cars and to focus on distinctive parts like treetops.



Multi-level optimization

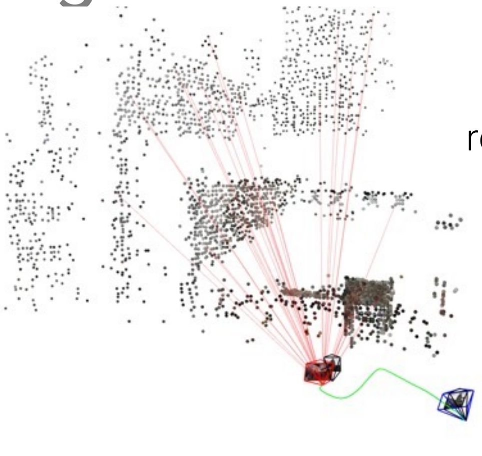


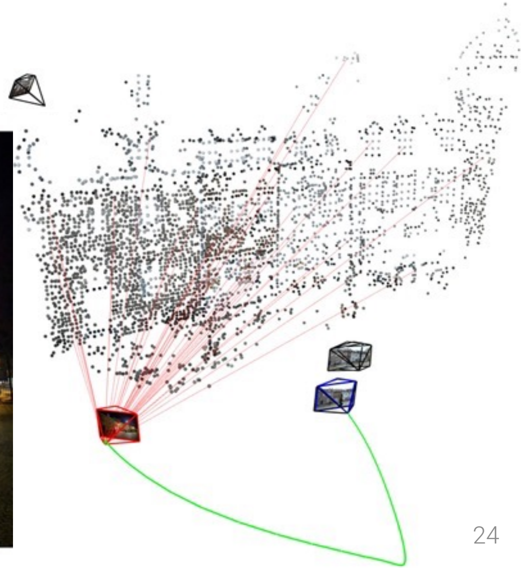
It can also ignore self-similarities that create local minima in the optimization, such as the brick wall shown here.



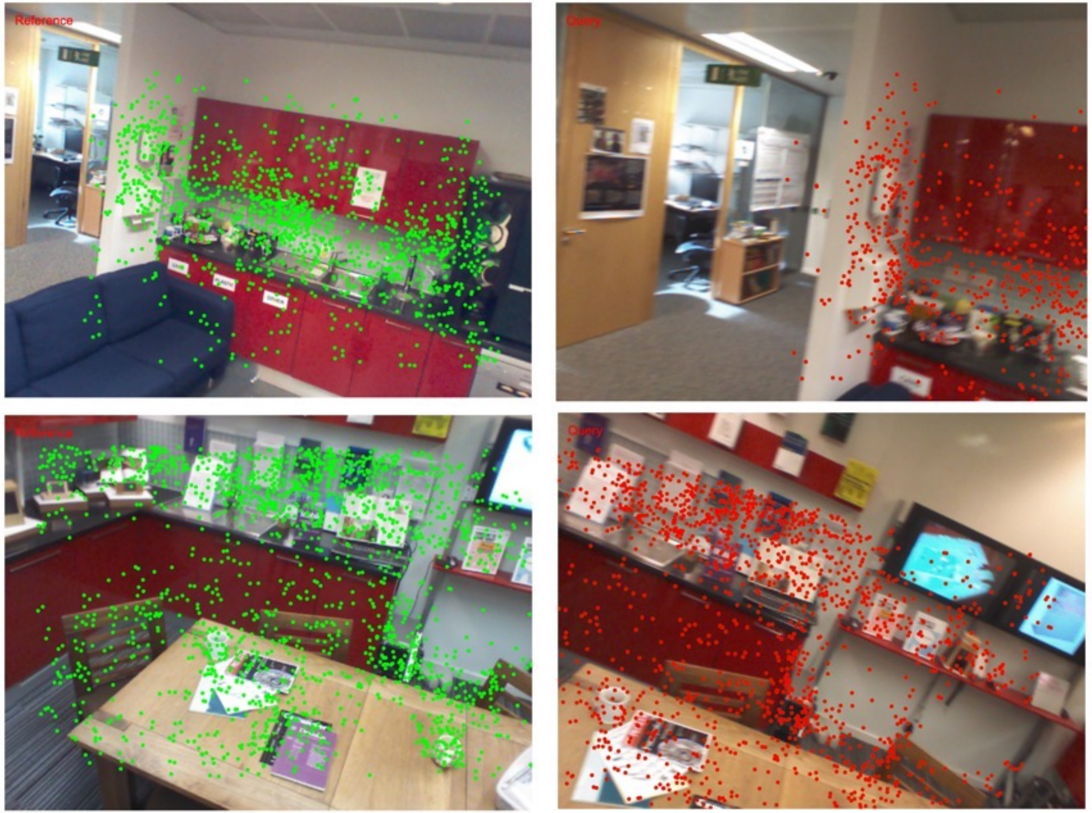
Large & robust convergence

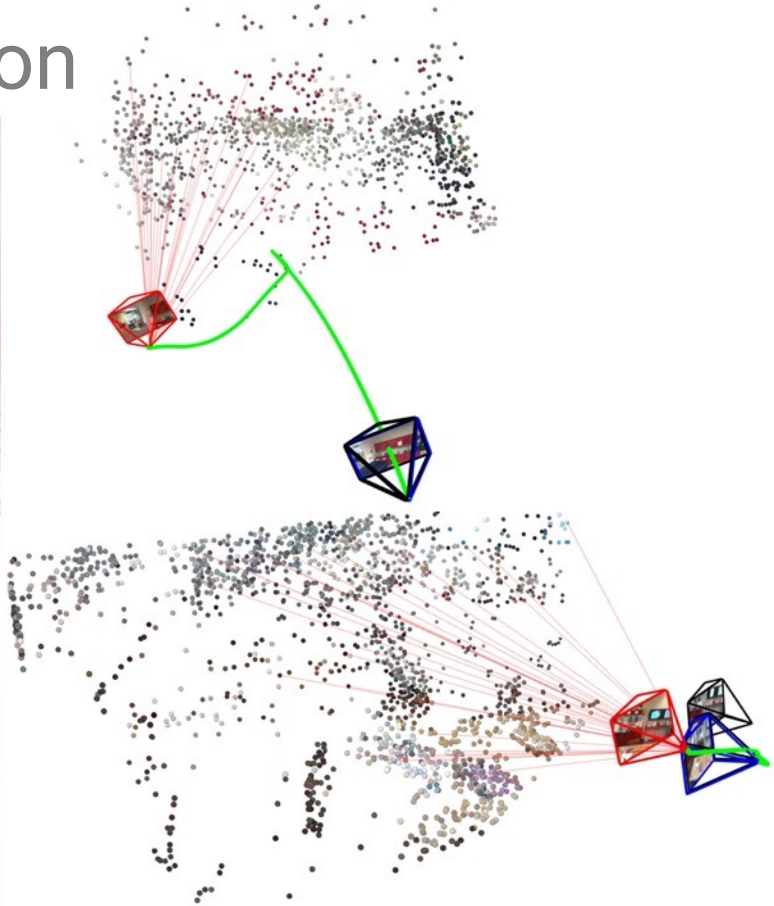
Direct alignment usually cannot handle large viewpoint or illumination changes. Differently, PixLoc has a large and robust basin of convergence thanks to the multilevel features.





Domain & scene generalization



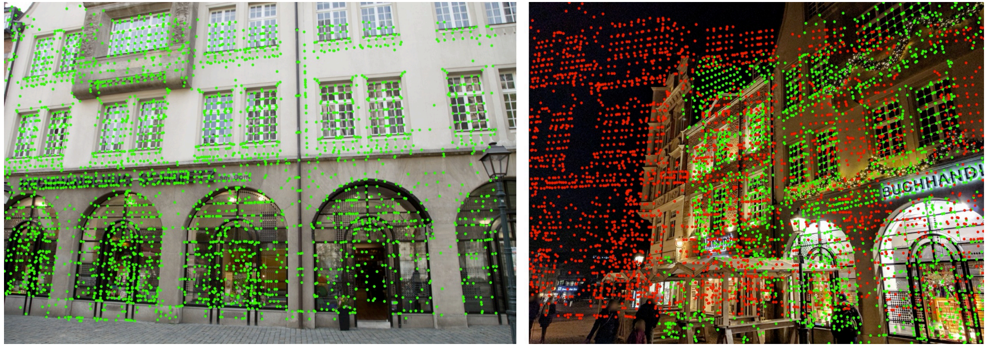


By learning only generic visual features, PixLoc generalizes well across environments.
A model trained only on outdoor scenes works well with indoor data that has less texture and more motion blur.

BACK TO THE FEATURE

psarlin.com/pixloc

PixLoc = end-to-end pose estimation



learn **temporal priors**
from pose supervision only!