LEARNING PRIORS FOR VISUAL LOCALIZATION AND MAPPING

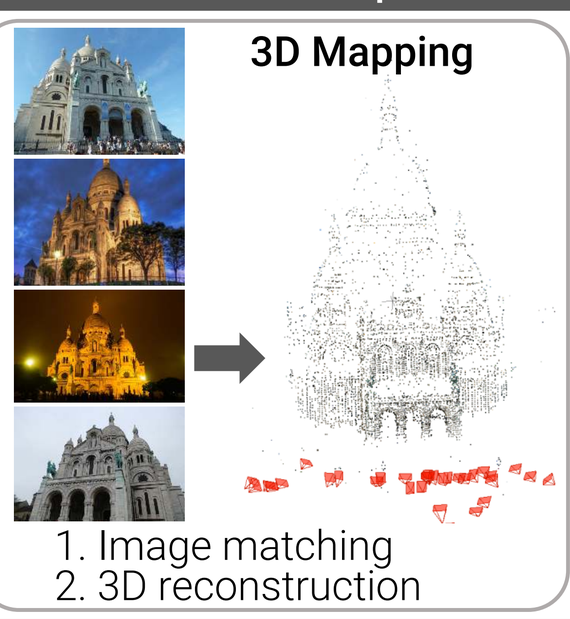
Paul-Edouard Sarlin – ETH Zurich – psarlin@ethz.ch

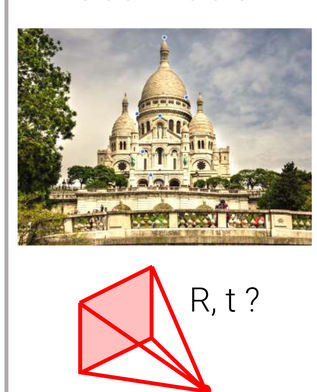
ETH zürich
ICVSS 2022
Sicily ~ 10-16 July
International Computer Vision Summer School

Abstract

Robots and Augmented Reality require **accurate visual positioning and 3D maps** that can cope with large areas, crowd-sourced data captured by various devices, and continual changes of the environment. We cover **three works that unlock these capabilities using the power of deep neural networks**: SuperGlue matches sparse image points with high robustness, Pixel-Perfect Structure-from-Motion yields more accurate maps from fewer images, and PixLoc learns to find stable objects useful for localization.

Problem setup & contributions





Challenges:

- Changing appearance
- Dynamic objects
- Large-scale environment
- High accuracy requirement
- Crowd-sourced data

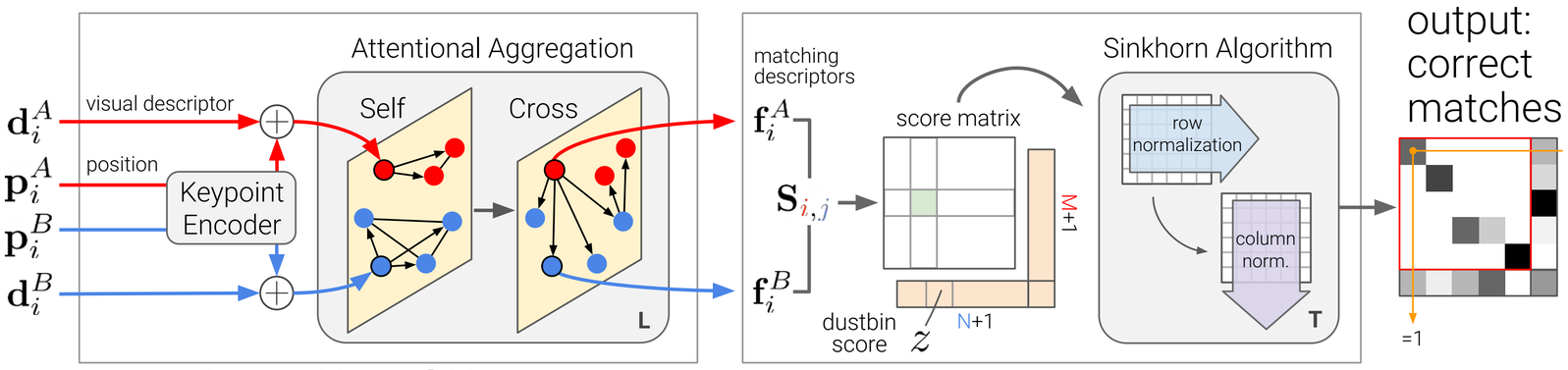
Our approach:
Learn high-level priors but not basic 3D geometry

1. SuperGlue: Learning Feature Matching with Graph Neural Networks [CVPR'20]

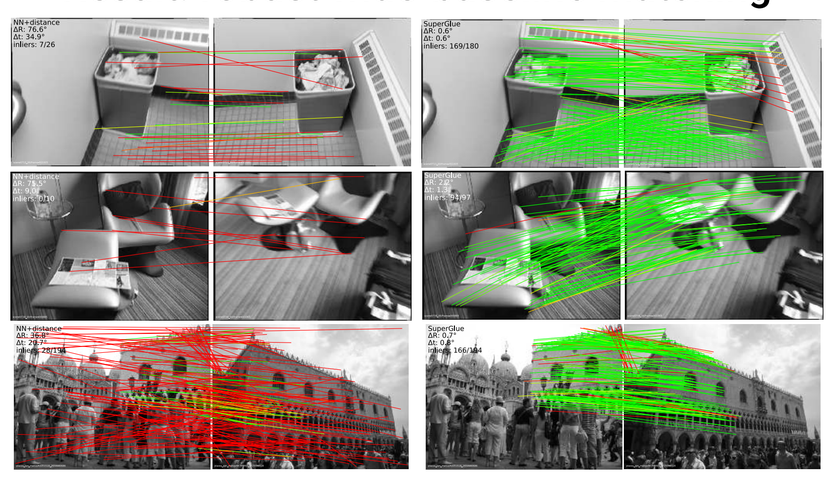
with Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich (Magic Leap)

Contribution: **a deep neural network for two-view image matching**

inputs: sets {●}, {●} of local features = keypoints + descriptors



Result: robust wide-baseline matching

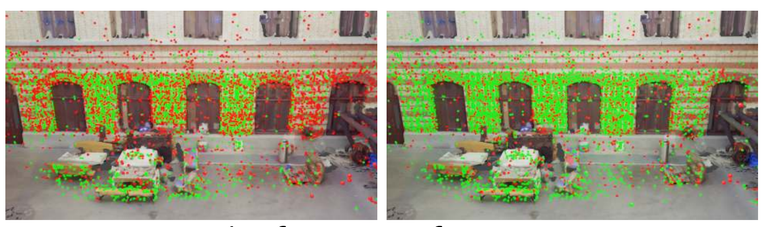


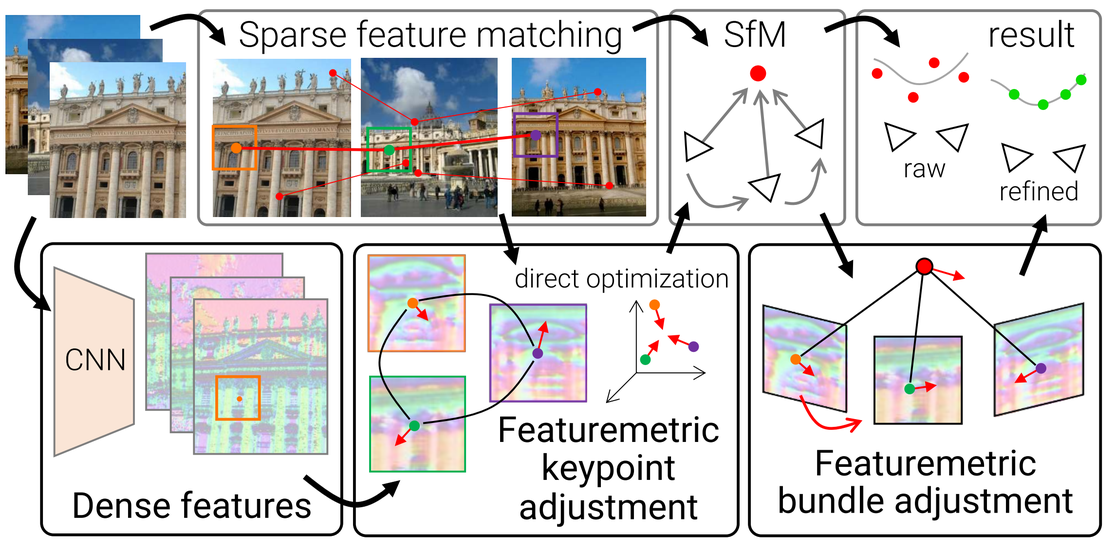
2. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement [ICCV'21]

with Philipp Lindenberger, Viktor Larsson, Marc Pollefeys (ETH Zurich, Microsoft)

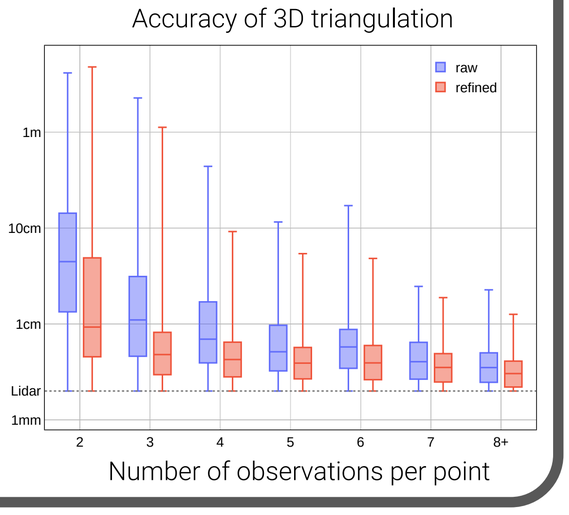
Contribution: **accurate SfM by aligning deep features across multiple views**

- Improves 3D points and camera poses
- Robust to real-world changes
- Scalable: use dense but local image info





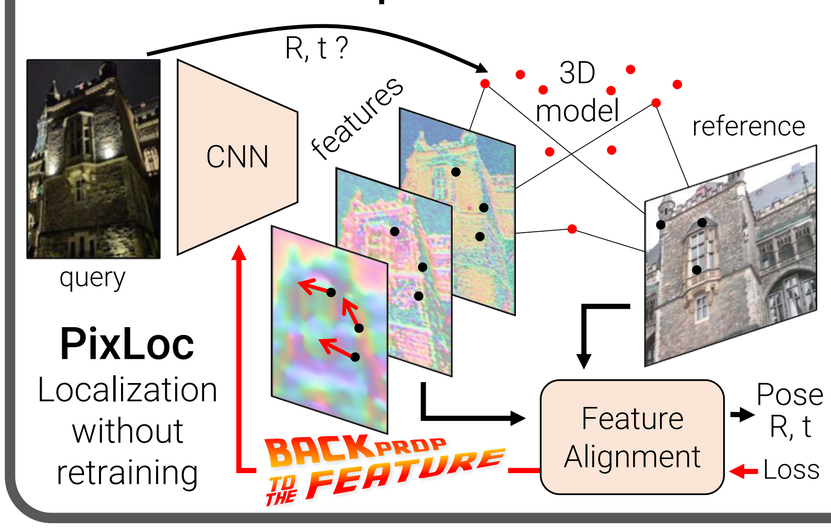
Result: more accurate 3D maps from fewer images



3. Back to the Feature: Learning Robust Camera Localization from Pixels to Pose [CVPR'21]

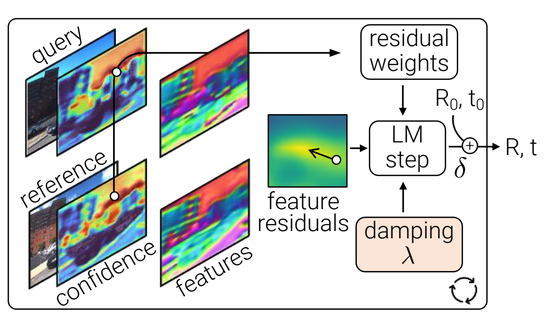
with A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, T. Sattler

Contribution: **a deep neural network that learns generic camera pose estimation end-to-end**



- Inputs: query image with coarse initial pose + 3D reference map (images + 3D points)
- Output: 6-DoF pose of the query

Learn to extract & align generic features
train once, deploy anywhere



Learns temporal priors
→ **ignore dynamic objects**