LIFT: Learned Invariant Feature Transform

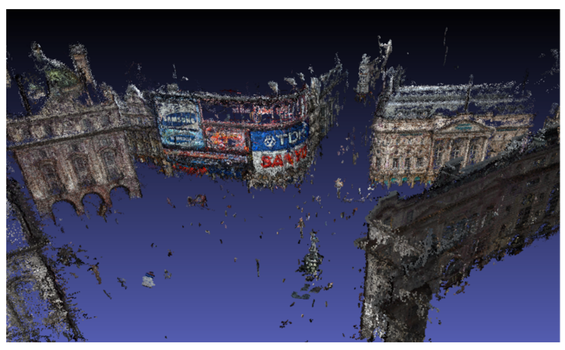
Kwang Moo Yi*,[1], Eduard Trulls*,[1], Vincent Lepetit[2], and Pascal Fua[1]
[1] École Polytechnique Fédérale de Lausanne [2] Graz University of Technology (* Equal contribution)
kwang.yi@epfl.ch, eduard.trulls@epfl.ch, lepetit@icg.tugraz.at, pascal.fua@epfl.ch

Motivation & Contributions

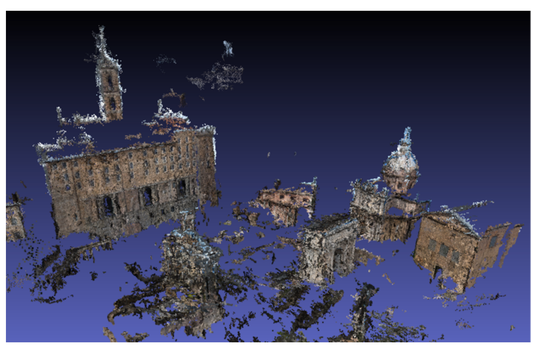
- We propose an **integrated, fully-differentiable deep network,** for keypoint detection, orientation estimation and feature description.
- **Joint optimization** improves overall performance.
- **Outperforms the state-of-the-art** on multiple datasets.
- Provides an off-the-shelf **replacement for SIFT,** with a practical computational time: **1.5x-3x** that of SIFT.
- **Code is available:** https://github.com/cvlab-epfl/LIFT

Training with Patches

- Train with **patches** to make the problem tractable and scalable.
- Two **SfM datasets:** Piccadilly ('pic') and Roman Forum ('rf').
- Keep only SfM points, i.e. we learn to find **repeatable points.**



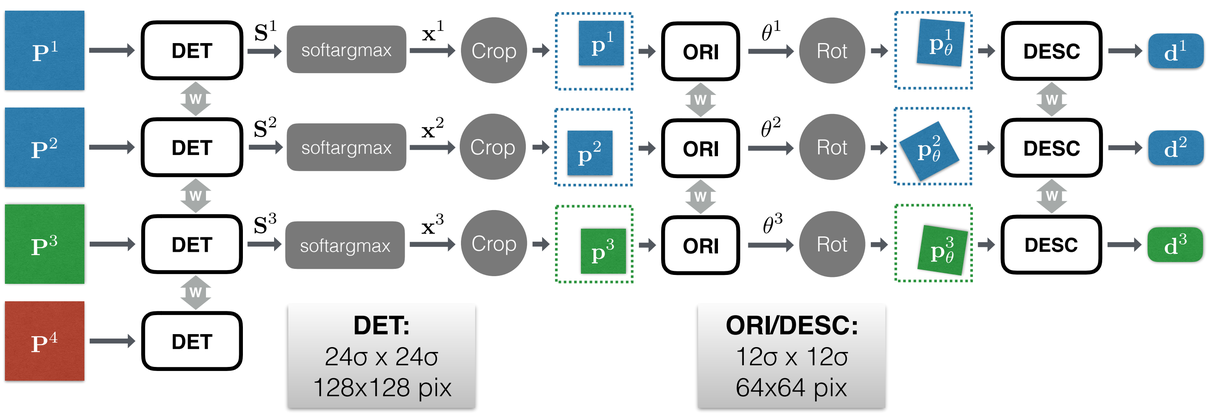
Piccadilly



Roman Forum

Quadruplet Siamese Network

- Training patches on **SIFT locations,** perturbed to avoid biases.
- **Quadruplet:** training the full pipeline requires non-keypoints, matching keypoints, and non-matching keypoints.



A Single Cost Function

We can optimize jointly with a single global loss:

$$\min_{\{f_\mu, g_\phi, h_\rho\}} \underbrace{\mathcal{L}_{\text{class}}(\mathbf{P}^1, \mathbf{P}^2, \mathbf{P}^3, \mathbf{P}^4)}_{\text{detector, orientation, descriptor}} + \gamma \underbrace{\mathcal{L}_{\text{pairwise}}(\mathbf{P}^1, \mathbf{P}^2, \mathbf{P}^3)}_{\text{detector, descriptor}}$$

$$\sum_{i=1}^{4} \alpha_i \max\left(0, \left(1 - \text{softmax}\left(f_\mu\left(\mathbf{P}^i\right)\right) y_i\right)\right)^2$$

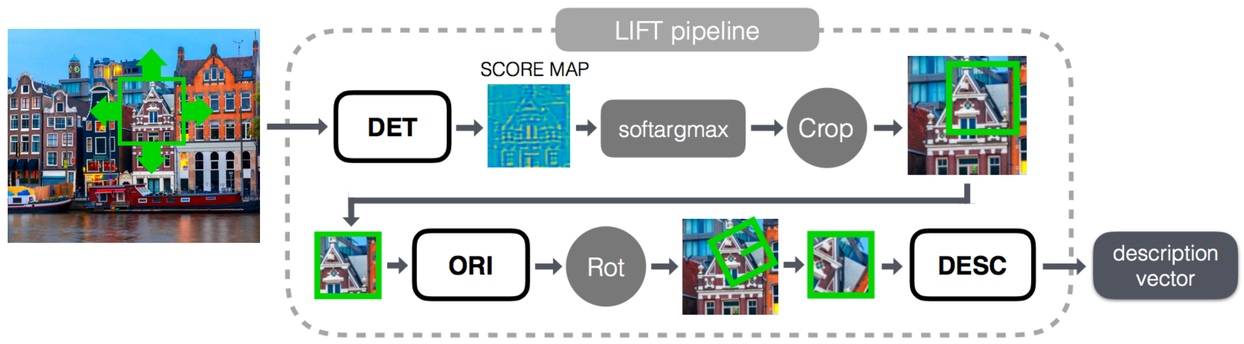
$$\left\| h_\rho\left(G\left(\mathbf{P}^1, \text{softargmax}\left(f_\mu\left(\mathbf{P}^1\right)\right)\right)\right) - h_\rho\left(G\left(\mathbf{P}^2, \text{softargmax}\left(f_\mu\left(\mathbf{P}^2\right)\right)\right)\right) \right\|_2$$

$$+ \max\left(0, C - \left\| h_\rho\left(G\left(\mathbf{P}^1, \text{softargmax}\left(f_\mu\left(\mathbf{P}^1\right)\right)\right)\right) - h_\rho\left(G\left(\mathbf{P}^3, \text{softargmax}\left(f_\mu\left(\mathbf{P}^3\right)\right)\right)\right) \right\|_2\right)$$

$$G\left(\mathbf{P}, \mathbf{x}\right) = \text{Rot}\left(\mathbf{P}, \mathbf{x}, g_\phi\left(\text{Crop}\left(\mathbf{P}, \mathbf{x}\right)\right)\right)$$

Spatial Transformers (Rot/Crop) are used as differentiable tools for image transformations. Note that these modules are not trained.

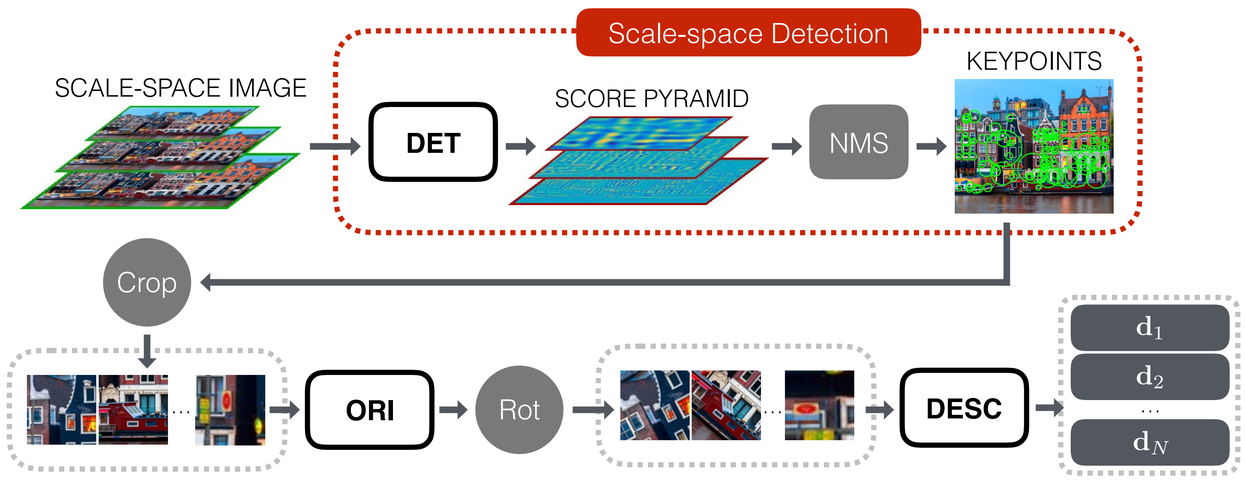
Integrated LIFT Network



- **DET, ORI, DESC:** Based on state-of-the-art deep networks.
- **Differentiable "Glue":** Spatial Transformers & softargmax.

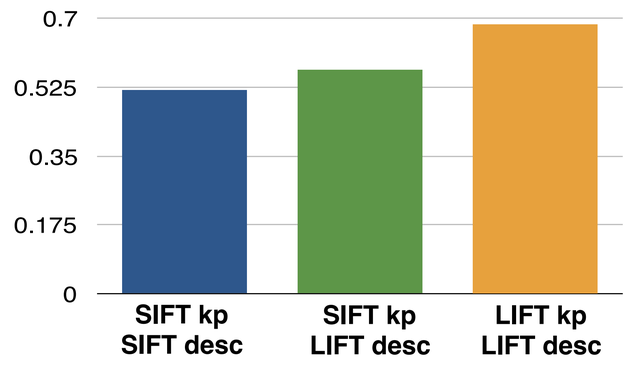
Run-time Pipeline

Detector runs in **scale-space with Non-Maximum Suppression.** The Orientation Estimator and Descriptor only process keypoints.



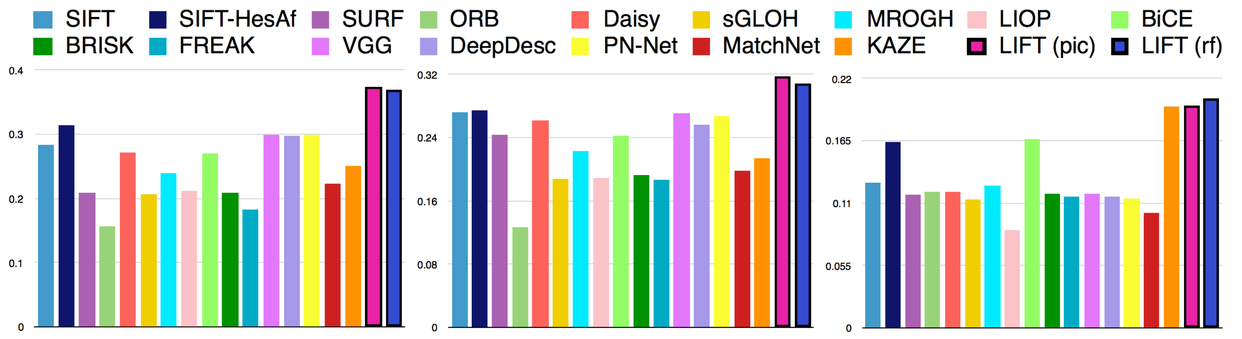
Joint Optimization

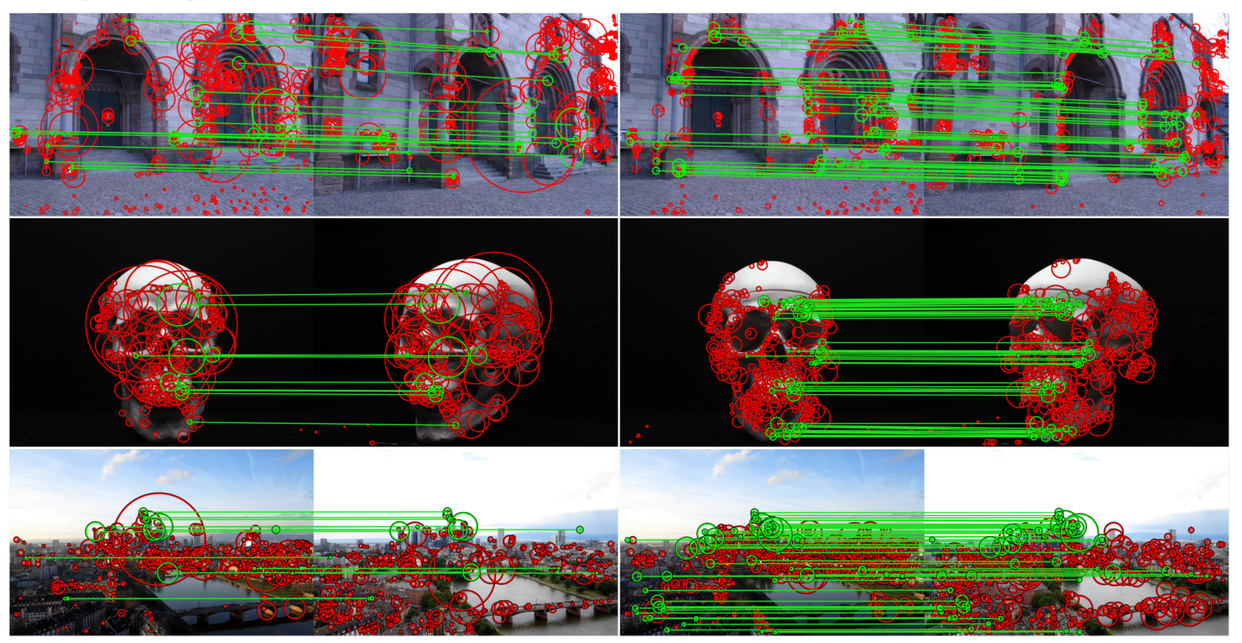
- **Descriptor performance,** in terms of NN mAP.
- **LIFT** descriptor works best with **LIFT** keypoints.
- Joint optimization is **key**.



Evaluation

- **Datasets:** 'Strecha', 'DTU', 'Webcam'.
  - 'Strecha': wide-baseline stereo (urban scenes).
  - 'DTU': viewpoint changes (objects).
  - 'Webcam': natural illumination changes, same viewpoint (outdoor).
- **Metric:** Matching score to capture full-pipeline performance.
  - The ratio of correct matches recovered in the shared viewpoint region.
- **Results:** best performance on all datasets, with 'rf' and 'pic'.





SIFT (still #3 overall) Ours