Motivation
Most learned features evaluated on patch pair classification task measuring false positive rate at 95% true positive rate (FPPR95) [3]
- Do better FPPR95 scores translate to better matching performance? What is the impact of typical ranking steps? (e.g., Lowe’s ratio test and mutual nearest neighbor constraint to avoid ambiguous matches, geometric verification to prune outliers requires good precision for manageable runtimes)
- More matches between similar images do not necessarily imply a better performance under extreme illumination and viewpoint changes. How well do learned features perform under such conditions?

Overview
Matching local image features is a key task in computer vision. For more than a decade, hand-crafted features such as SIFT have been used for this task. Recently, new features learned from data have been proposed and shown to improve on SIFT in terms of discriminative power. This work is dedicated to an extensive experimental evaluation of local features in a practical setting.

Benchmark
Single evaluation protocol to benchmark local image feature performance in a practical setting:
- Raw image-to-image matching performance
  (under e.g., blur, exposure, day-night, scale, rotation, planar, internet, etc.)
- Image-based reconstruction performance
  (measuring impact of local feature matching performance on Bag-of-Words image retrieval, Structure-from-Motion, and Multi-View Stereo)

Methods
- SIFT: RootSIFT [1]
- SIFT-PCA: RootSIFT with PCA projection [4]
- DISP-SIFT: Domain-size pooled SIFT [5]
- ConvOpt: Learned descriptor using convex optim. [8]
- TFeat: Shallow learned descriptor [7]
- LIFT: Learned keypoint detector and descriptor [6]

Reconstruction Performance
- Evaluation using Structure-from-Motion and Multi-View Stereo
- Exhaustive image matching for Fountain (11 images), HerzJesu (8 images), South Building (128 images), Madrid Metropolis (124 images), Gendarmenmarkt (146 images)
- Image retrieval with matching against top-100 retrieved images for Alamo (2015 images), Roman Forum (234 images), Cornell (2014 images)

Insights
- Patch classification performance does not translate to more complex image-based reconstruction task
- Previous image-based reconstruction datasets too easy as a benchmark (Fountain, Herzjesu, South Building)
- Learned features better than RootSIFT but not better than advanced hand-crafted features still better
- Learned features exhibit strong variation in performance for different datasets
- Significant room for improvement, especially in the hard cases where all methods fail (e.g., day-night)