Learning Feature Representations for Localization and Mapping

Mihai Dusmanu

Johannes Schönberger, Marc Pollefeys

Ignacio Rocco, Tomas Pajdla, Josef Sivic, Akihiko Torii, Torsten Sattler

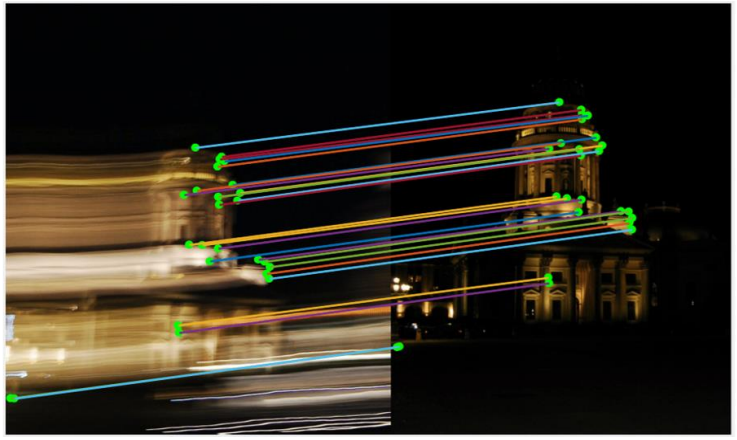
1. An introduction to local features
2. D2-Net – detect-and-describe approach to local features
3. Open research question
4. Multi-view keypoint refinement for accurate reconstructions

Structure-from-Motion Revisited, Schönberger et al., CVPR 2016

Why do we need local features?

SfM, SLAM, Visual Localization, AR…

Efficiency / scalability



What do we want from local features?

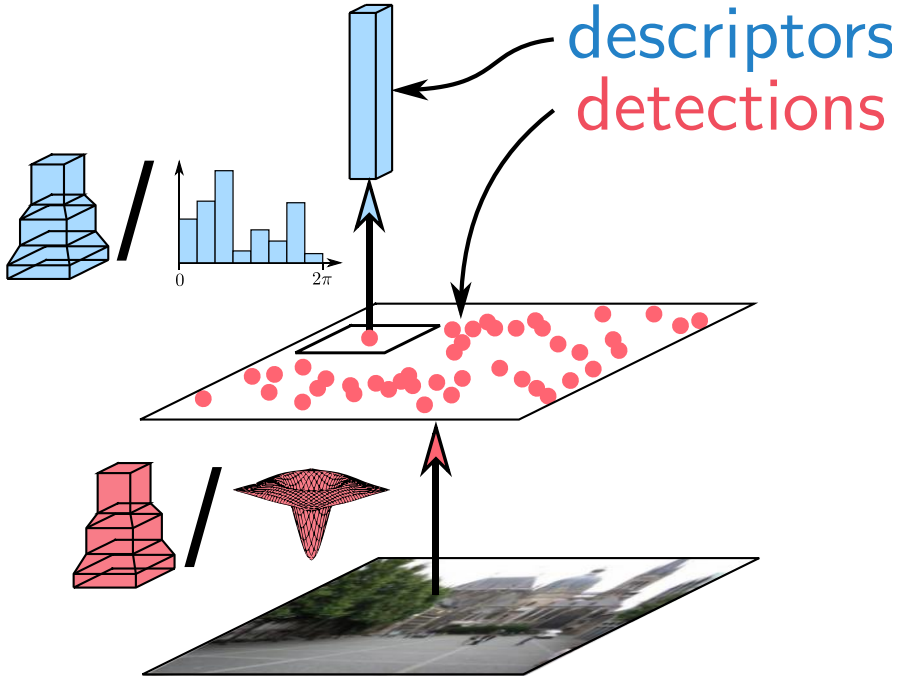
Repeatability and matchability

Robustness (viewpoint / seasonal / day-night changes, motion blur)

Detect-Then-Describe

Classic Approach

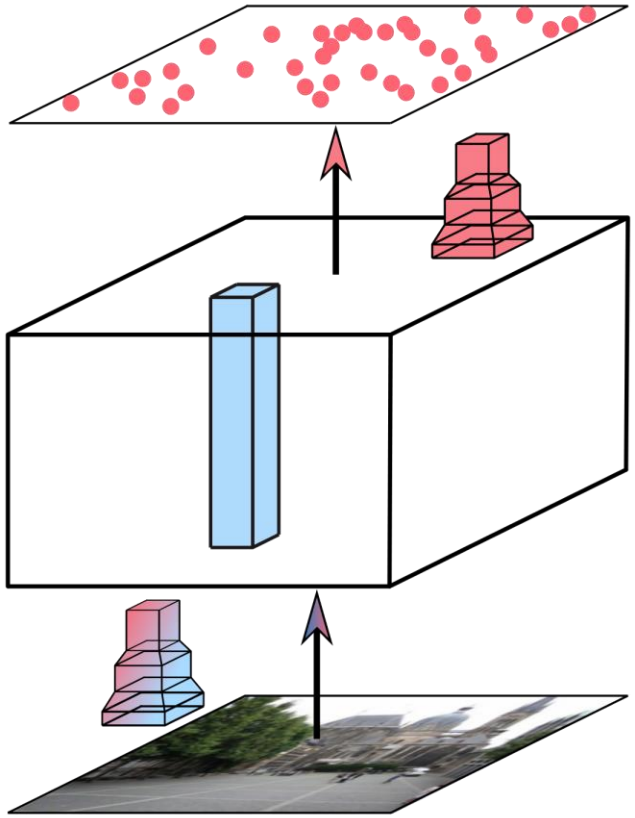
- Detectors:
  - Handcrafted: DoG, Harris, Hessian, …
  - Trainable: TILDE, TCDET, Quad-Networks, …
  - Hybrid: HesAffNet
- Descriptors:
  - SIFT, BRIEF, …
  - T-Feat, HardNet, GeoDesc, …
- Full pipeline:
  - SIFT, ORB, …
  - LIFT, LF-Net, …



Describe-Then-Detect

DELF

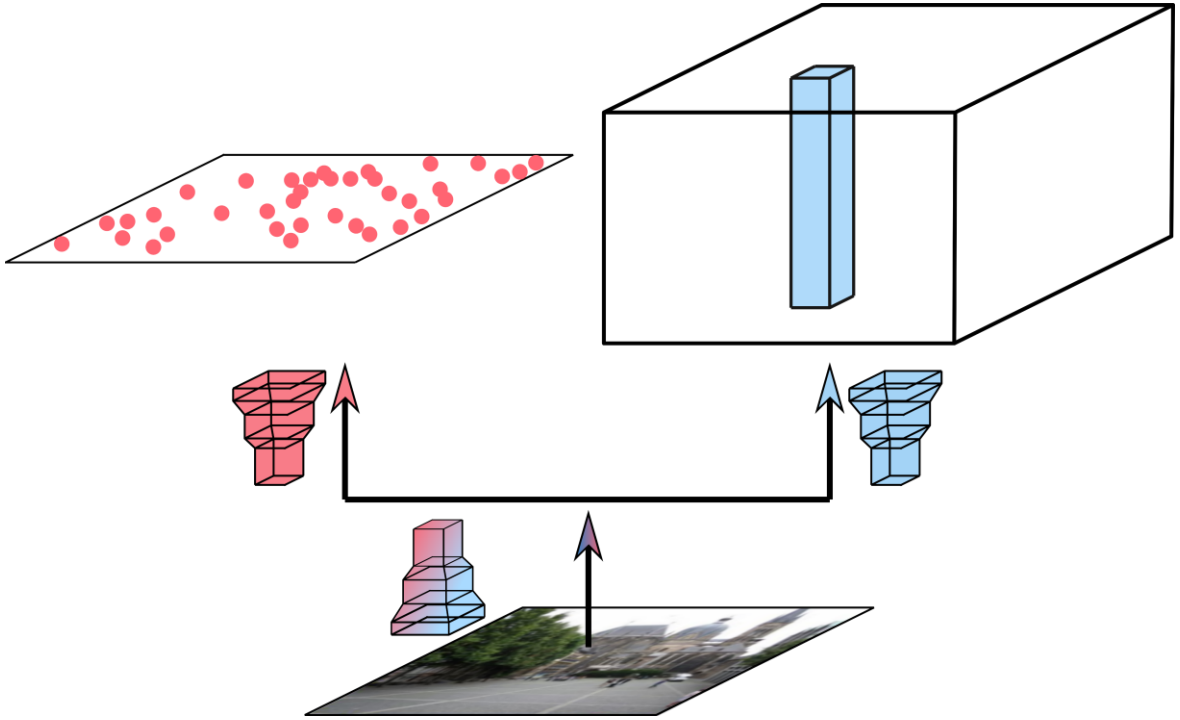
Large-Scale Image Retrieval with Attentive Deep Local Features
Noh et al., ICCV 2017



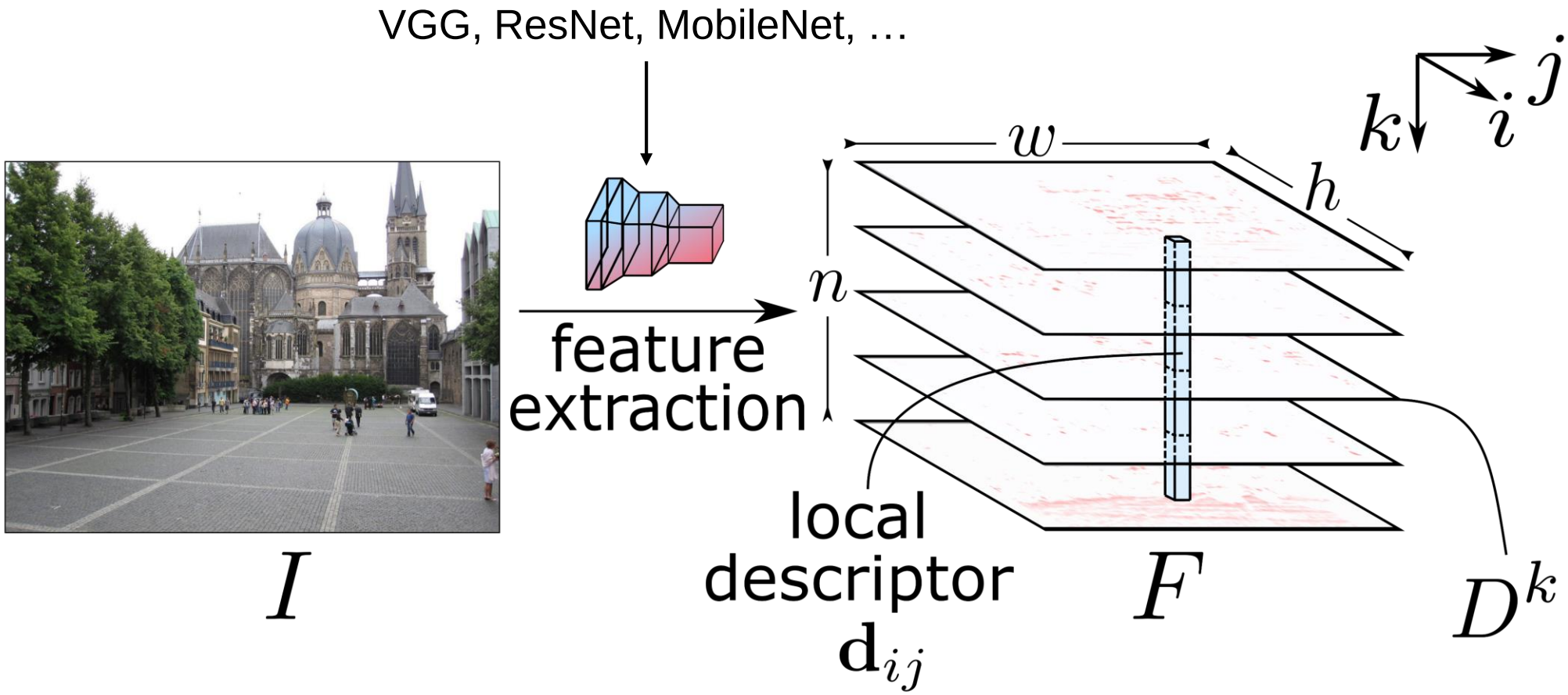
Shared Encoder

SuperPoint

SuperPoint: Self-Supervised Interest Point Detection and Description
DeTone et al., CVPR Workshops 2018



D2-Net – Detect-and-Describe

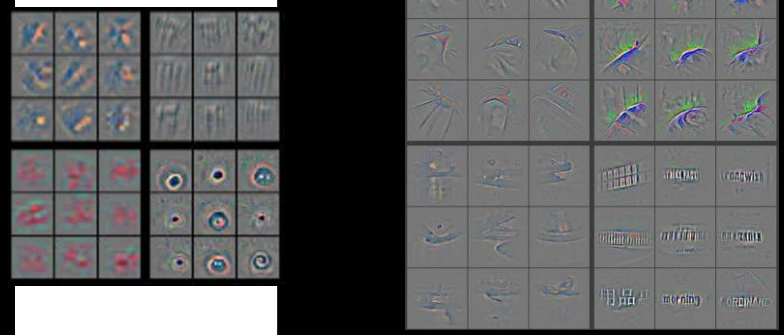


D2-Net – What layer to choose?

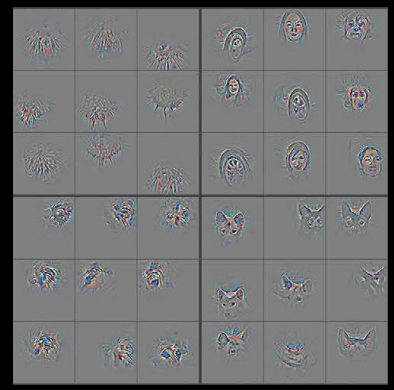
Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

Low-Level Features

Mid-Level Features

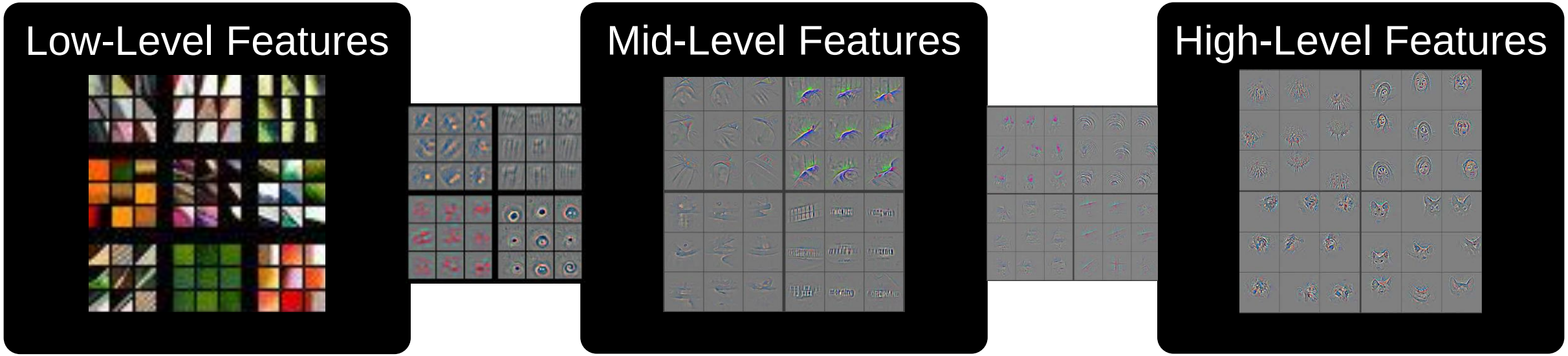


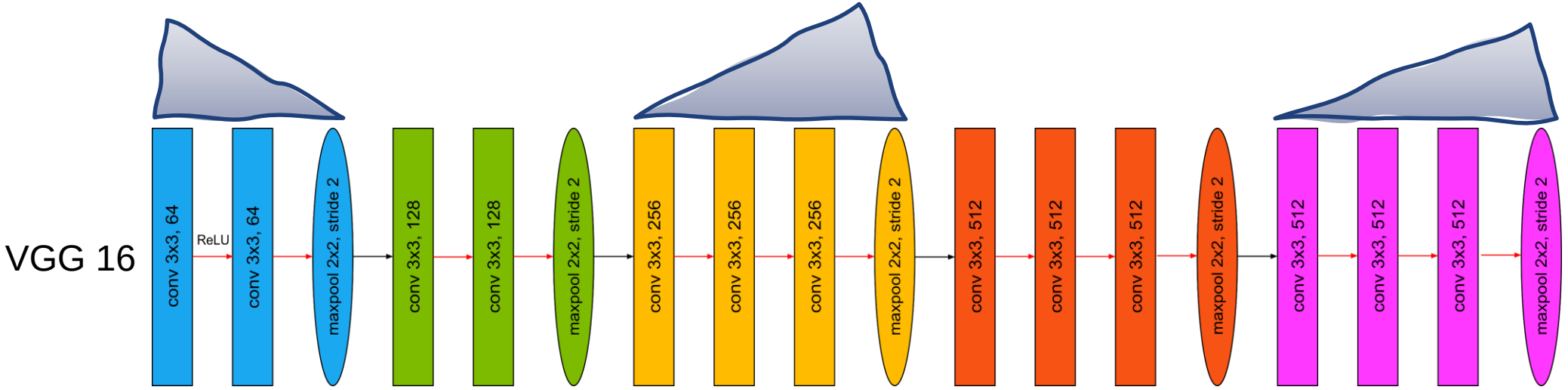
High-Level Features



D2-Net – What layer to choose?

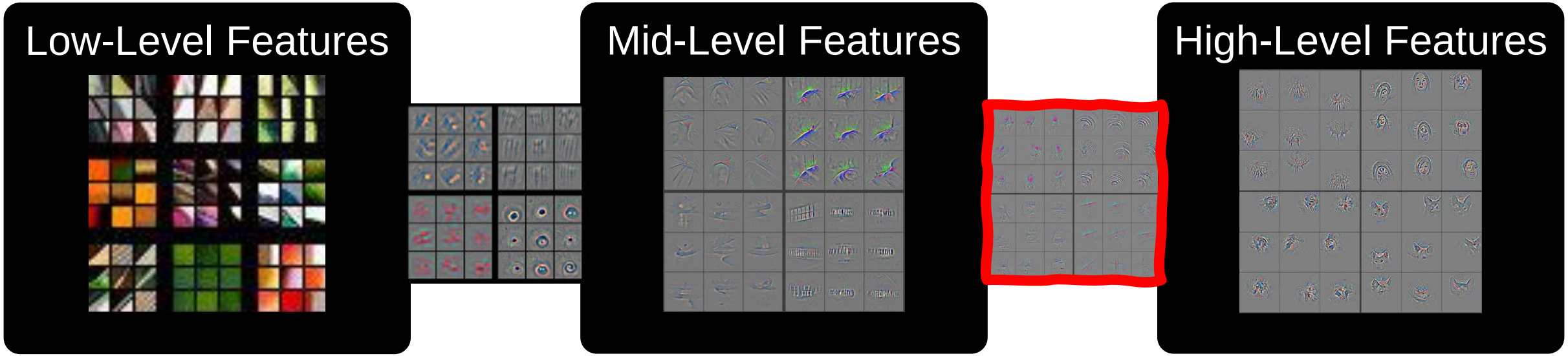
Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014

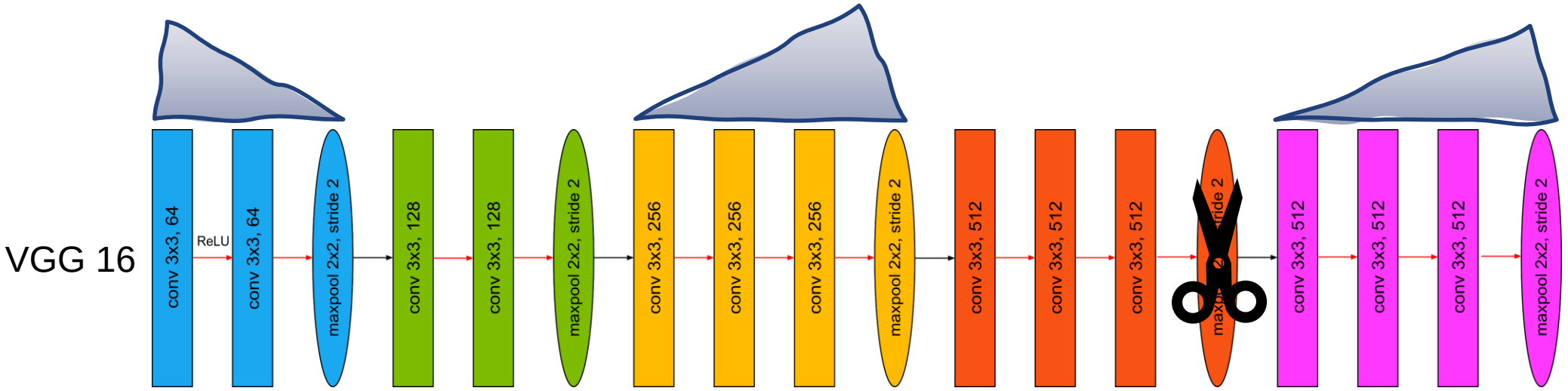




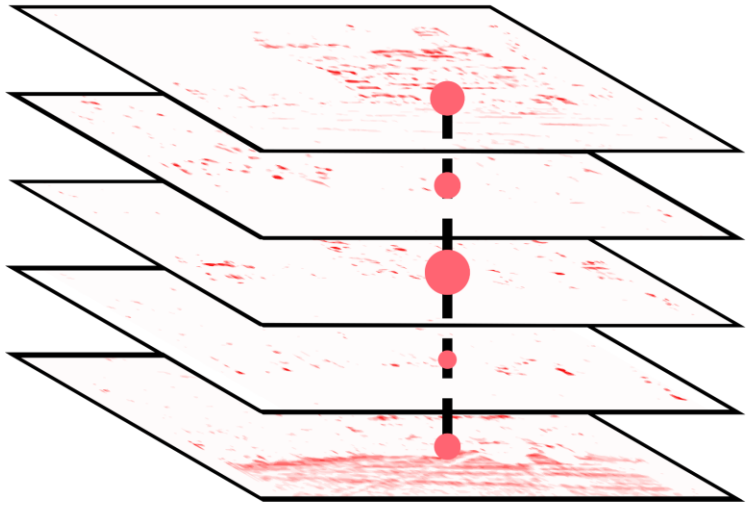
D2-Net – What layer to choose?

Visualizing and Understanding Convolutional Networks, Zeiler & Fergus, ECCV 2014



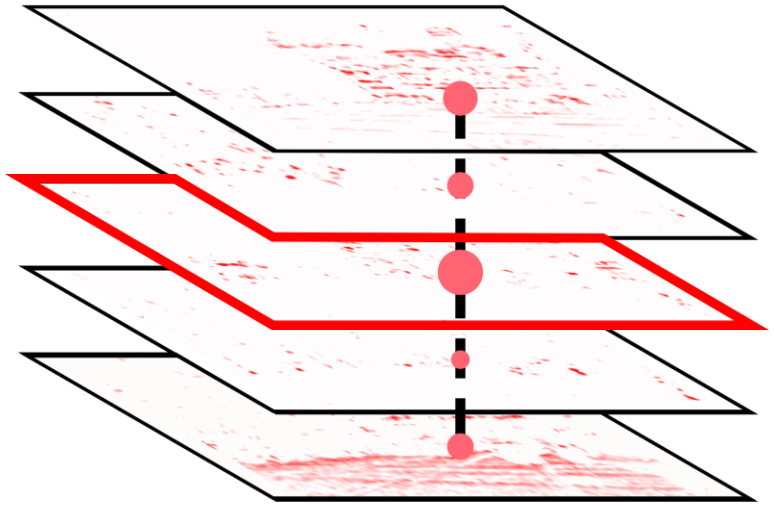


D2-Net – Keypoint Detection



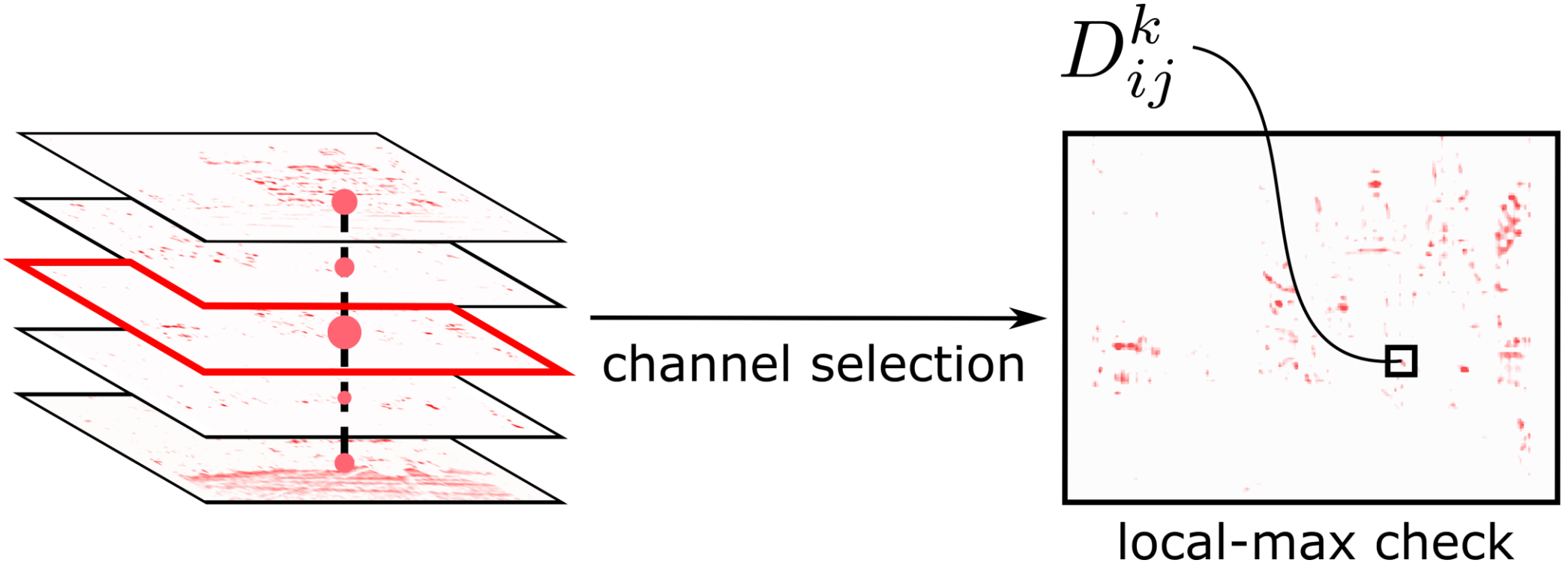
$$(i, j) \text{ is a detection} \iff D_{ij}^{k} \text{ is a local max. in } D^{k} \quad ,$$
$$\text{with } k = \arg\max_{t} D_{ij}^{t} \quad .$$

D2-Net – Keypoint Detection



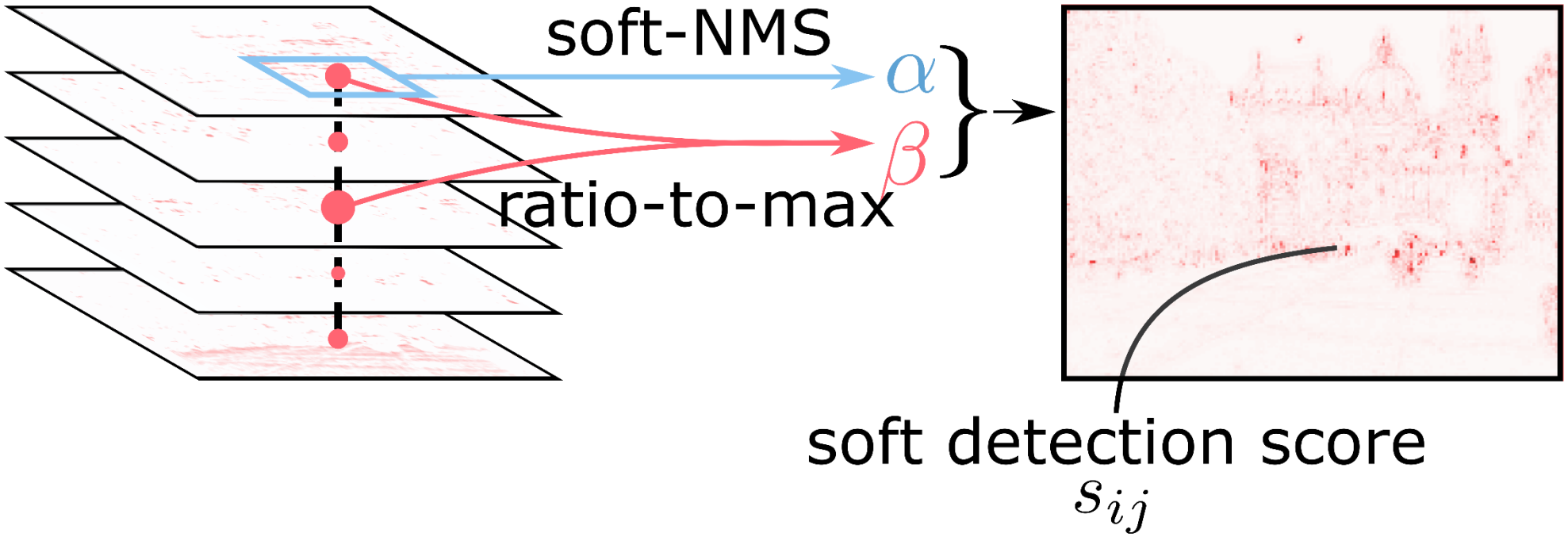
$$(i, j) \text{ is a detection} \iff D_{ij}^{k} \text{ is a local max. in } D^{k} \quad ,$$
$$\text{with } k = \arg\max_{t} D_{ij}^{t} \quad .$$

D2-Net – Keypoint Detection



$$(i, j) \text{ is a detection} \iff D_{ij}^k \text{ is a local max. in } D^k \quad ,$$
$$\text{with } k = \arg\max_t D_{ij}^t \quad .$$

D2-Net – Soft Keypoint Detection for Training



$$\beta_{ij}^k = D_{ij}^k \Big/ \max_t D_{ij}^t \qquad \alpha_{ij}^k = \frac{\exp\left(D_{ij}^k\right)}{\sum_{(i',j')\in\mathcal{N}(i,j)} \exp\left(D_{i'j'}^k\right)} \qquad s_{ij} \propto \max_k \left(\alpha_{ij}^k \beta_{ij}^k\right)$$

D2-Net – Joint Detection-Description Loss

- Triplet loss for description

$$m(c) = \max\left(0, M + p(c)^2 - n(c)^2\right)$$

  - Negative sample: in-image-pair negative mining - filter out repetitive structures
- Weighted average of triplet losses over all correspondences

$$\mathcal{L}(I_1, I_2) = \sum_{c \in \mathcal{C}} \frac{s_c^{(1)} s_c^{(2)}}{\sum_{q \in \mathcal{C}} s_q^{(1)} s_q^{(2)}} m(p(c), n(c))$$

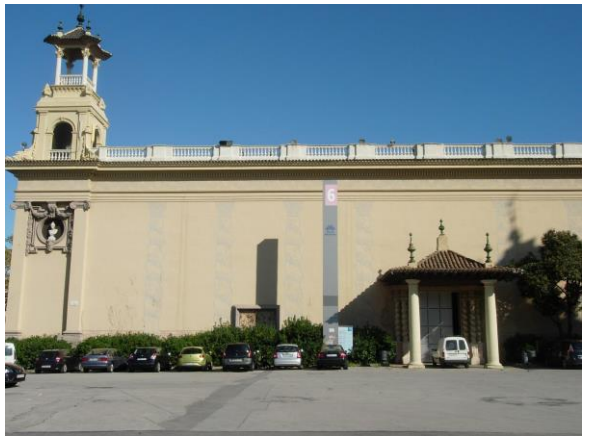
  - good correspondence ⇔ low triplet loss value
- Requires correspondences
  - MegaDepth: 196 different locations reconstructed with COLMAP SfM / MVS

D2-Net – Results

- Long-term Visual Localization Benchmark
  - https://www.visuallocalization.net
  - Different localization scenarios:
    - Different seasons / illumination conditions (including night-to-day)
    - Indoor localization
    - Autonomous driving
    - Suburban / Park scenes with vegetation
  - Ranked #1 on 3 datasets and #2 on 2 datasets
  - Using NetVLAD / VLAD retrieval

D2-Net – Summary

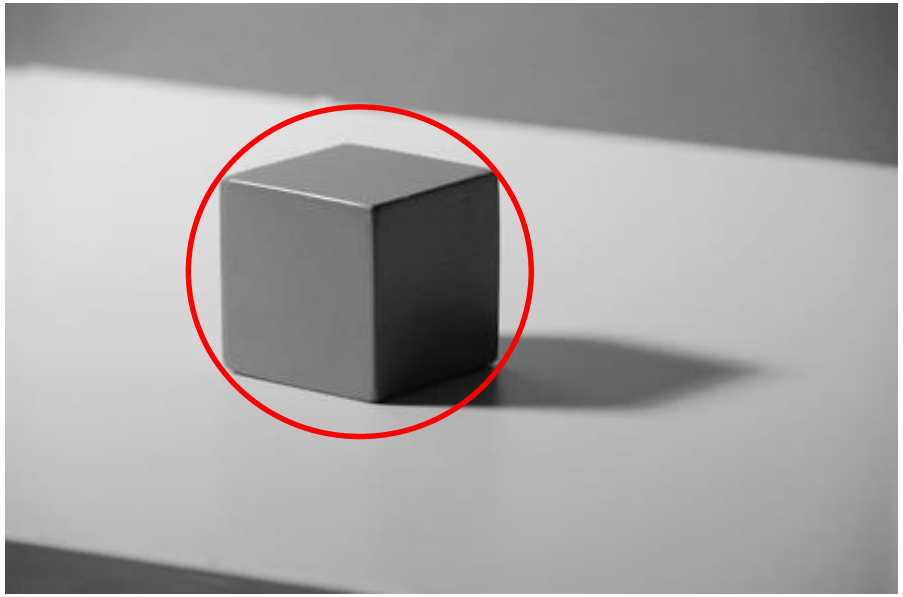
- Joint detection and description
- State-of-the-Art for local features on challenging camera localization tasks
- Versatile: not architecture-specific
- Problems:
  - Poor keypoint localization
    - Raw matching: beats SOTA starting at ~6px
    - >1 px reprojection error for 3D reconstructions
    - Large receptive field, max pooling
  - Feature ambiguity
    - Large receptive field

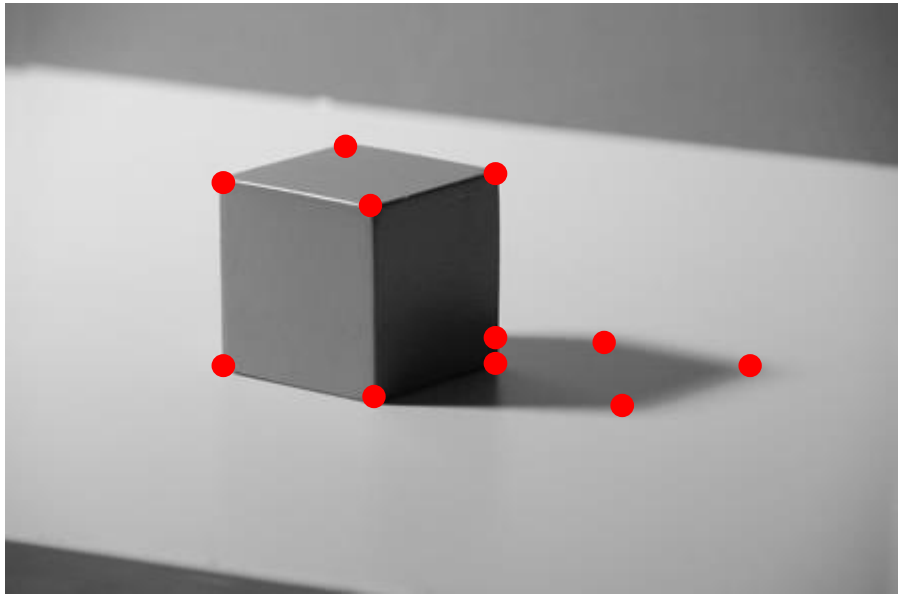


Will there be a Swiss-army-knife local feature ?

An extreme example…

- High level detections: robust but not well localized
- Low level detections: very well localized, but not as robust

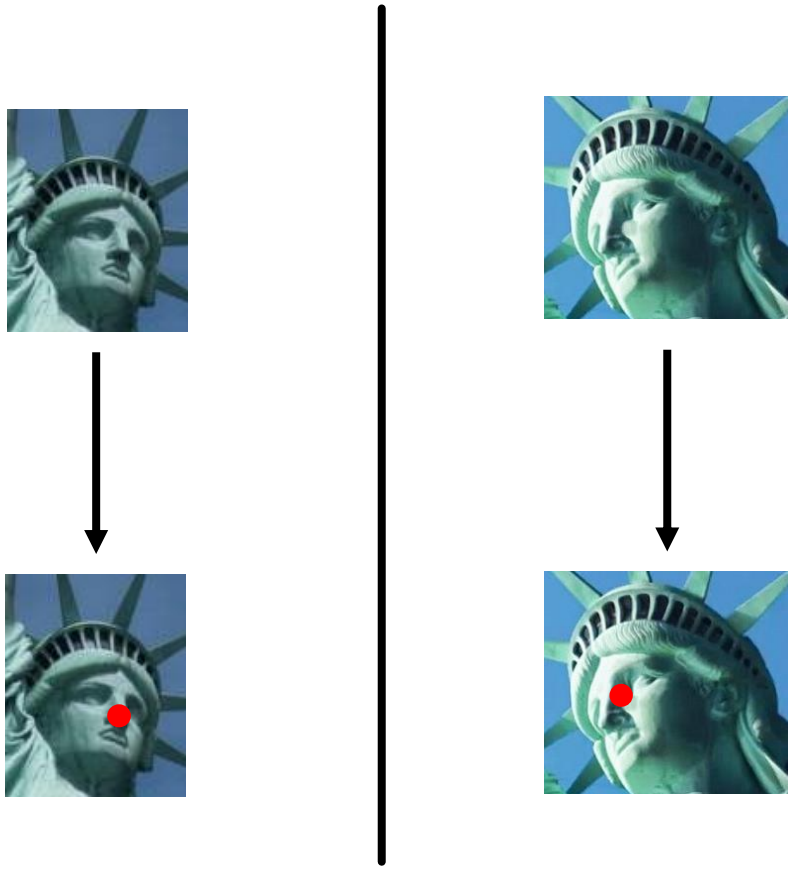




Brigham Young University – Idaho, The Department of Art, ART 110 – Beginning Drawing
https://courses.byui.edu/art110_new/art110/week06/Light_shadow.html

Feature extraction

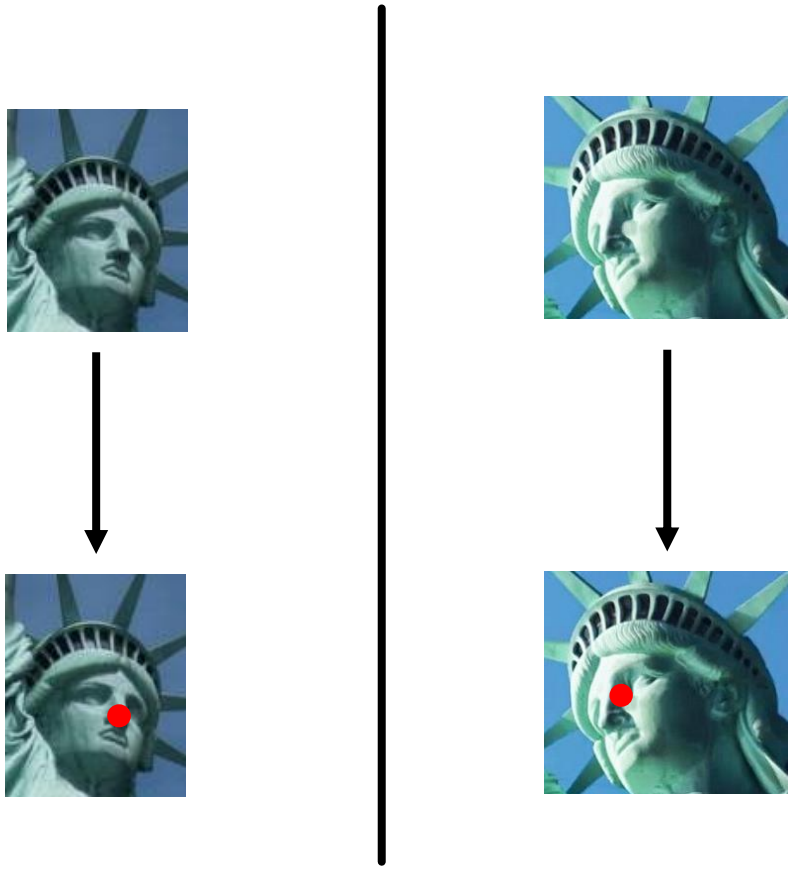
Independent for each image

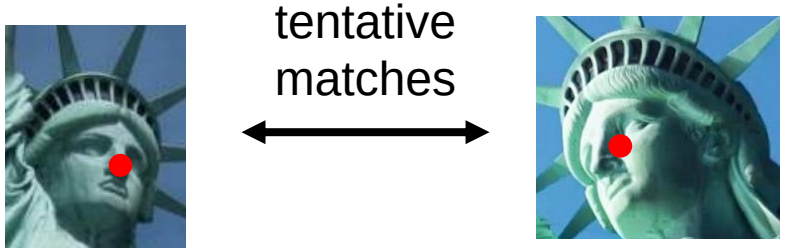


Microsoft

Feature extraction

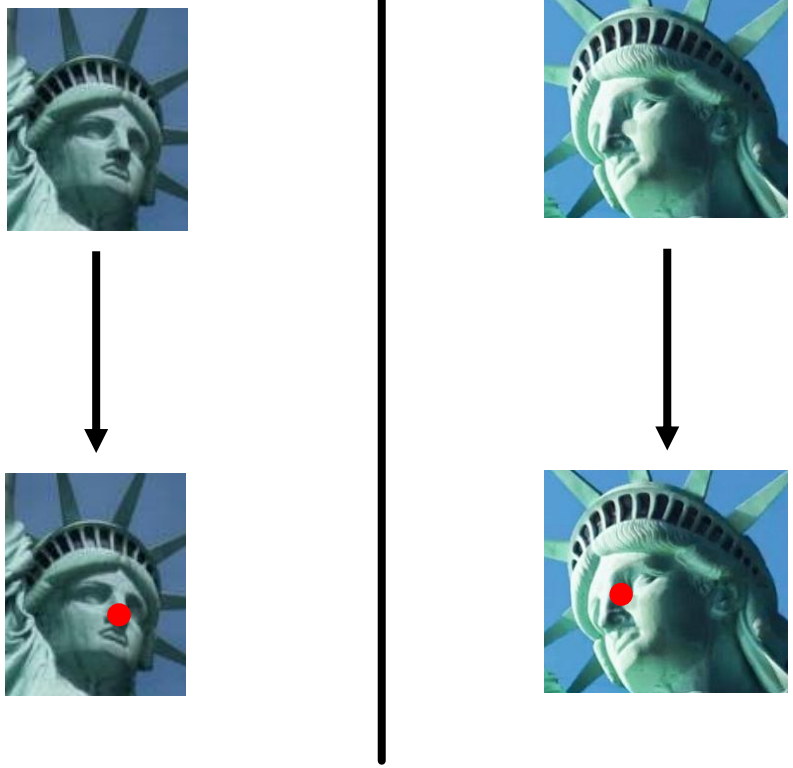
Independent for each image



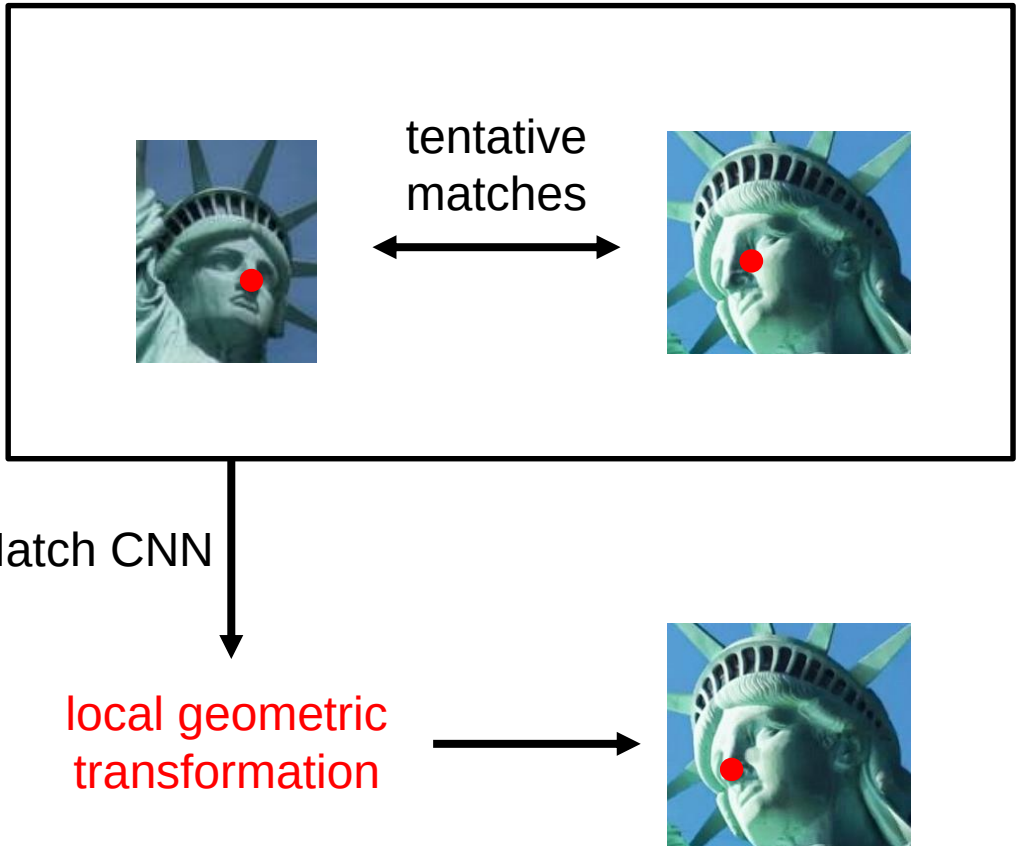


Feature extraction

Independent for each image



Patch matching



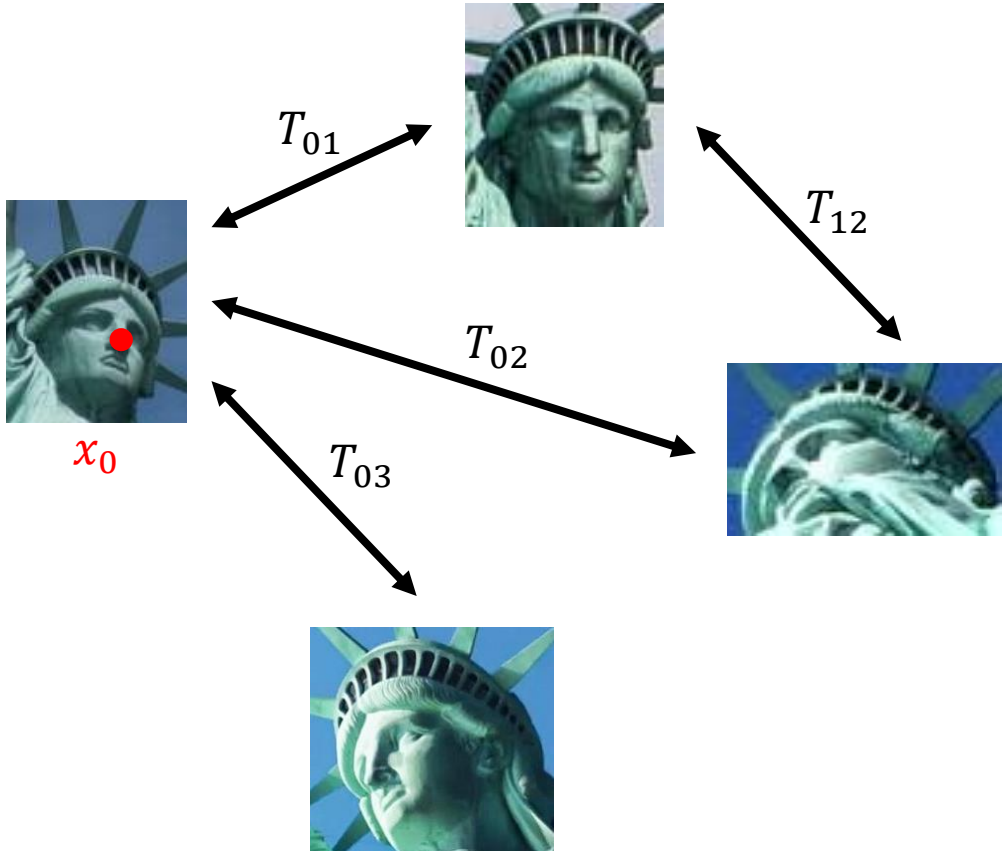
Refining with multiple views

Tentative matches graph

Challenges:

- Incorrect matches
- Inaccurate transformations
- (Very) large graphs
- Feature drift
- Repeated structures

$$\min_{x_k} \sum_{i \to j \text{ edge}} \rho\left(\|x_j - T_{ij} x_i\|^2\right)$$



Questions