Deep SFM - Proprosal

Rui Zhu, You-Yi Jau

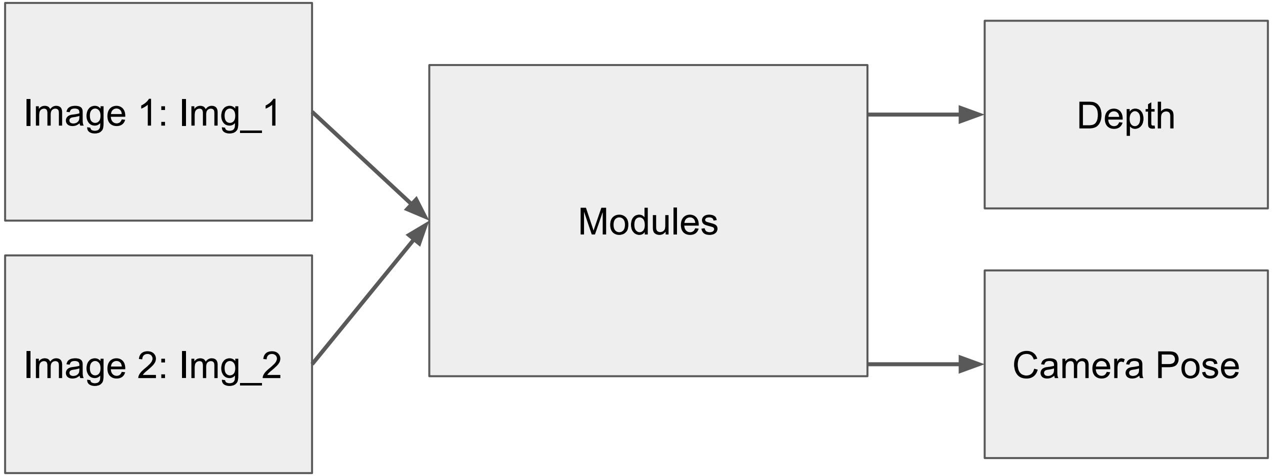
2018/11

Outlines

- Motivation
- Introduction
  - Traditional method
  - Baseline
- Proposed method
- References

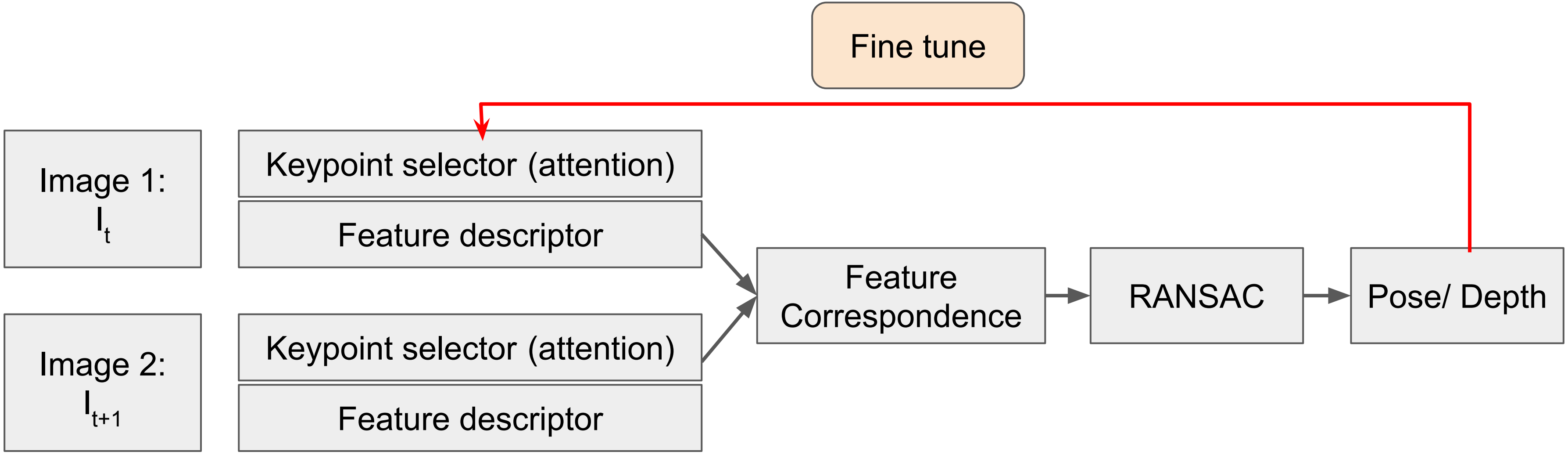
Motivation

- Modules are designed heuristically
- Submodules are optimized individually
- Can deep learning help?



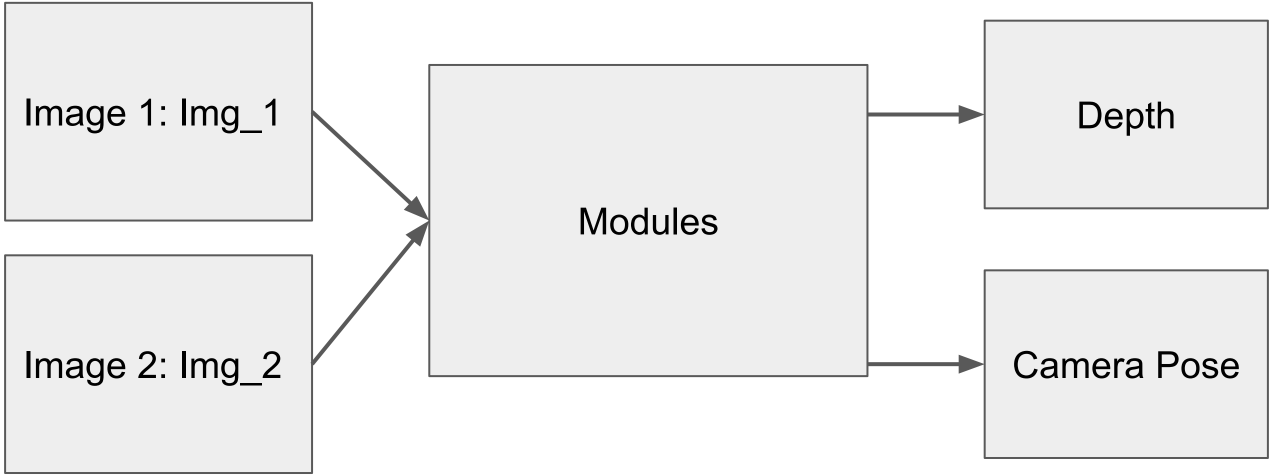
Conclusion: Keypoint-based method: Our method

- Goal
  - Optimize the modules based on output error (keypoint reprojection error)
  - **Self-supervised** training using estimated pose and depth.
  - Use temporal information to **adapt keypoint selector** (attention map)



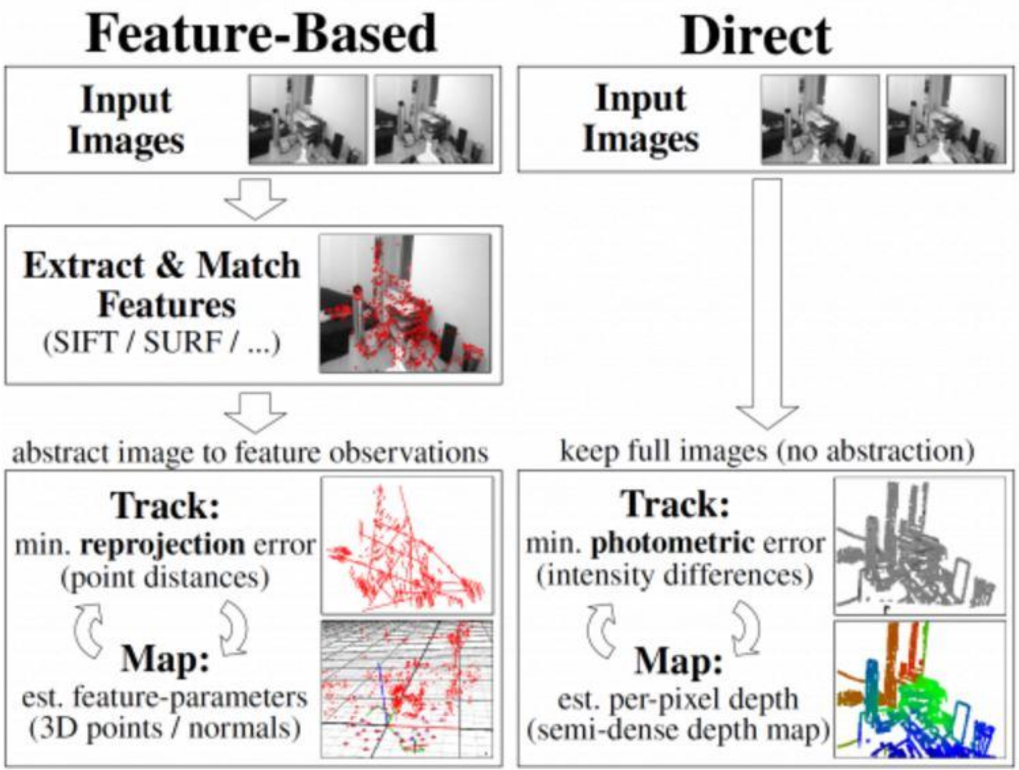
Pipeline

- Input/ Output
- Modules



How to design “Modules” with learning?

- Feature-based method
- Direct Method

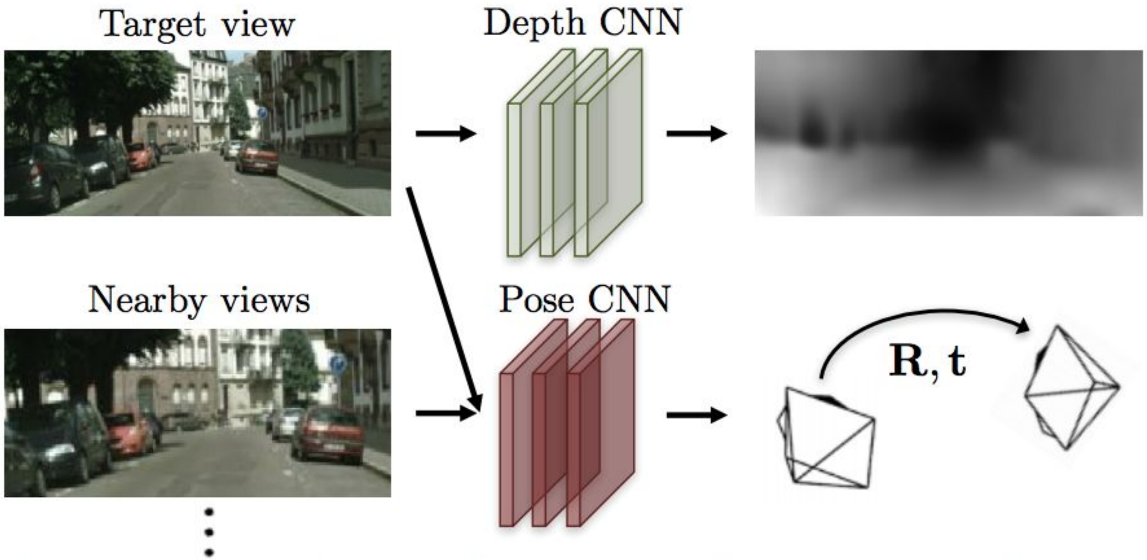


Medium: LSD-slam and ORB-slam2, a literature based explanation

Modules: Direct method

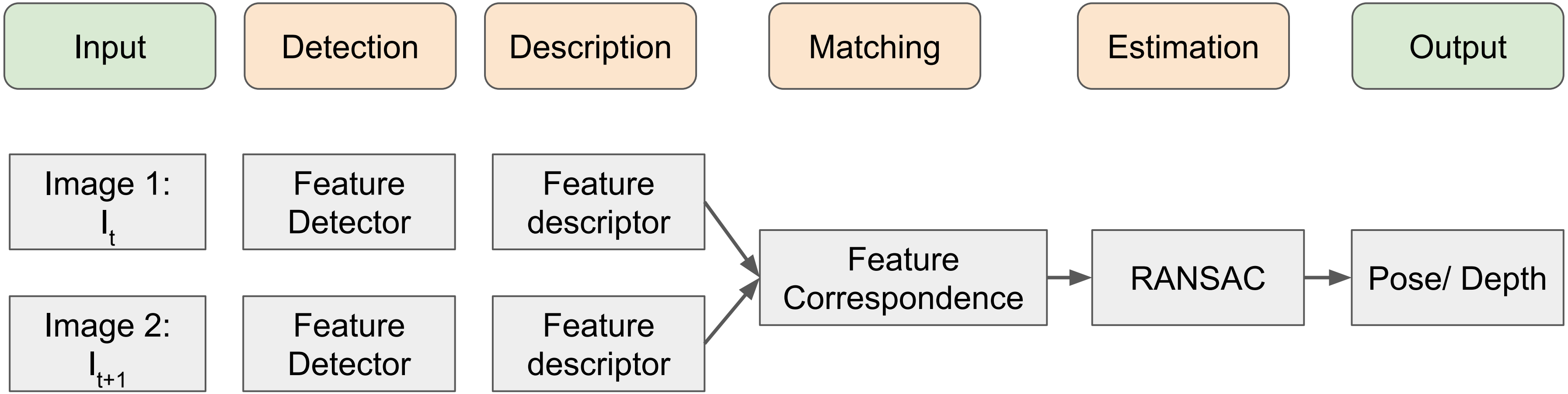
[Unsupervised learning of depth and ego-motion from video]

T Zhou, M Brown, N Snavely, DG Lowe - CVPR, 2017



(b) Testing: single-view depth and multi-view pose estimation.

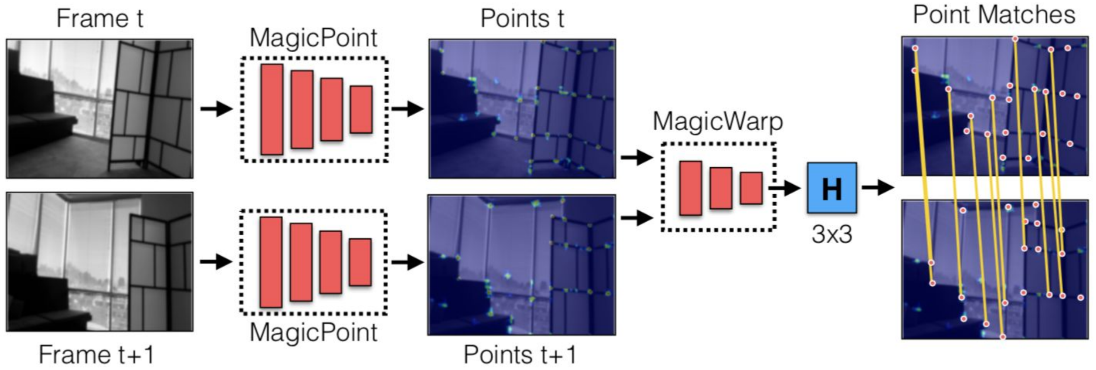
Modules: Keypoint-based method

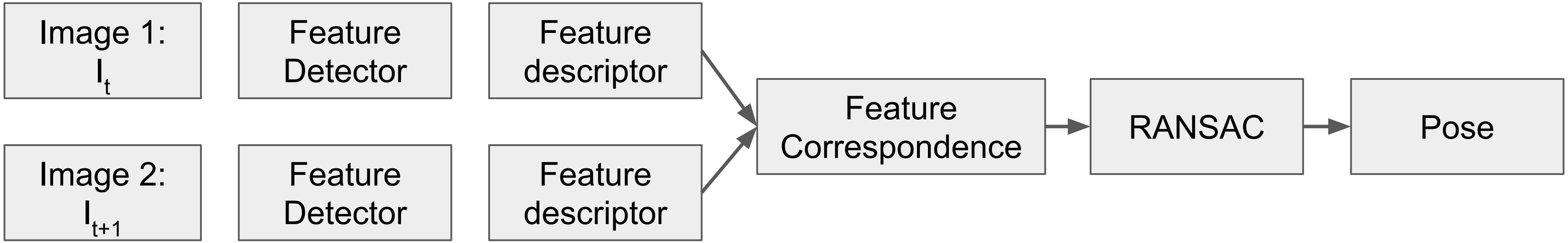


Keypoint-based method: Example

[Toward Geometric Deep SLAM]

D DeTone, T Malisiewicz, A Rabinovich - arXiv preprint arXiv:1707.07410, 2017





Keypoint-based method: Traditional method

Input | Detection | Description | Matching | Estimation | Output

- Fast Corner Detector
  - E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In ECCV, 2006.

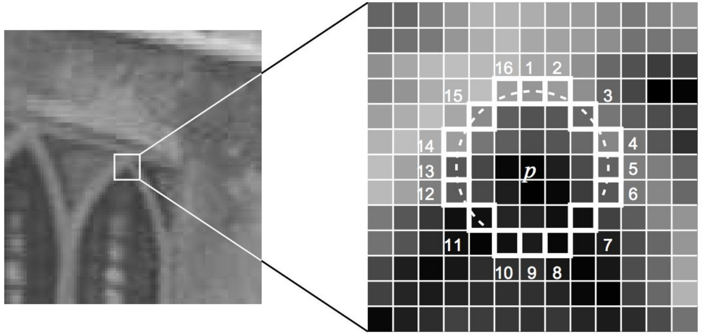
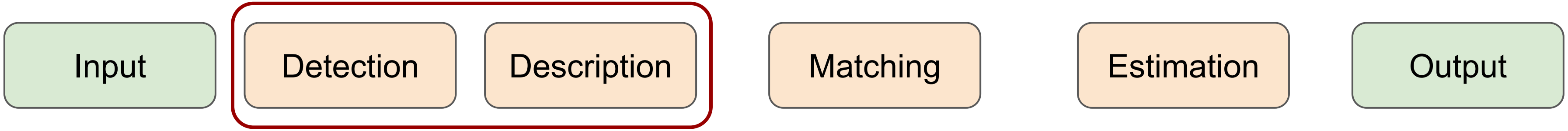


Figure 1. 12 point segment test corner detection in an image patch. The highlighted squares are the pixels used in the corner detection. The pixel at p is the centre of a candidate corner. The arc is indicated by the dashed line passes through 12 contiguous pixels which are brighter than p by more than the threshold.

Keypoint-based method: Traditional method



- Scale-Invariant Feature Transform (SIFT)
  - D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 2004.

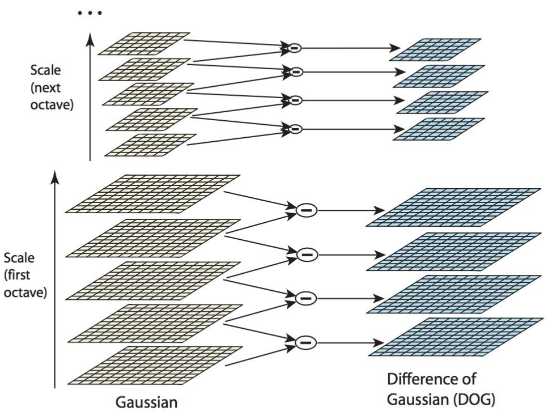


Figure 1: For each octave of scale space, the initial image is repeatedly convolved with Gaussians to produce the set of scale space images shown on the left. Adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images on the right. After each octave, the Gaussian image is down-sampled by a factor of 2, and the process repeated.

Keypoint-based method: Traditional method

Input | Detection | Description | **Matching** | Estimation | Output

- Nearest neighbor
    - SIFT: Euclidean distance

Keypoint-based method: Traditional method

Input | Detection | Description | Matching | **Estimation** | Output

- RANSAC
  - Homography matrix
    - [Motion and structure from motion in a piecewise planar environment]

      OD Faugeras, F Lustman - International Journal of Pattern Recognition and …, 1988
  - Essential Matrix
    - Eight-point algorithm

Keypoint-based method: Baseline

Input | Detection | Description | Matching | Estimation | Output

- Lift: Learned Invariant Feature Transform
  - Imitate the pipeline of SIFT

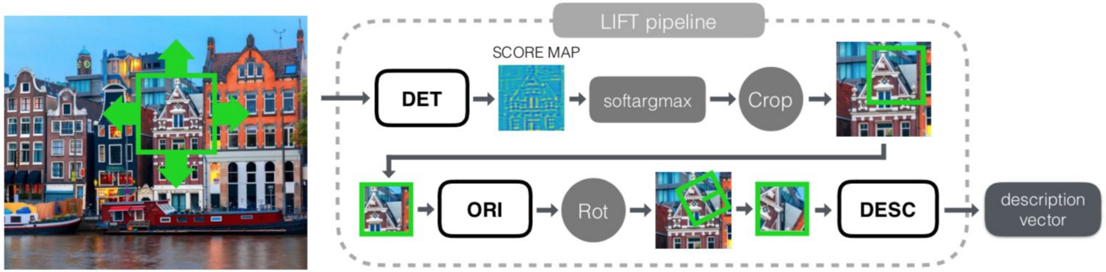


Fig. 1. Our integrated feature extraction pipeline. Our pipeline consists of three major components: the Detector, the Orientation Estimator, and the Descriptor. They are tied together with differentiable operations to preserve end-to-end differentiability.[1]

[Lift: Learned invariant feature transform]() KM Yi, E Trulls, V Lepetit, P Fua - European Conference on Computer Vision, 2016

Keypoint-based method: Baseline

Input | Detection | Description | Matching | Estimation | Output

- SuperPoint: Self-Supervised Interest Point Detection and Description

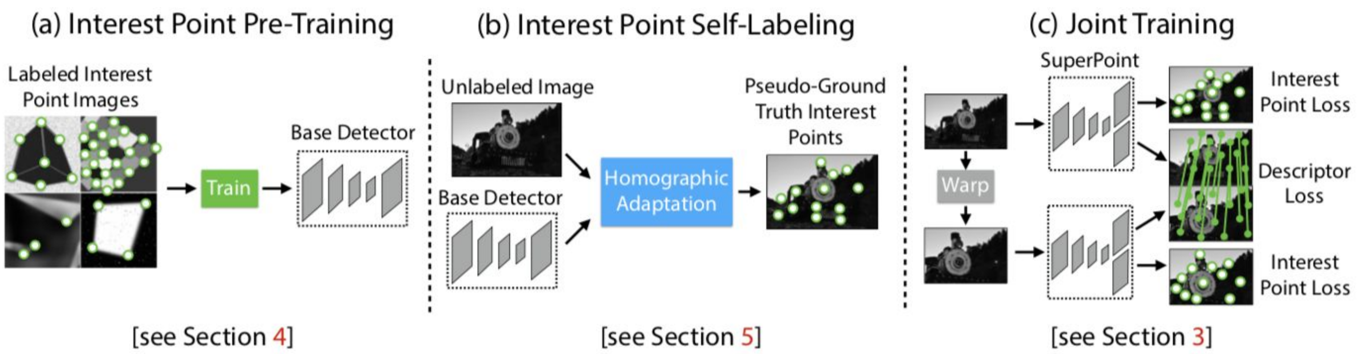
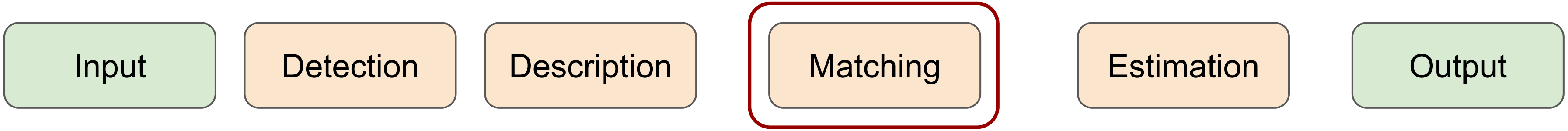


Figure 2. **Self-Supervised Training Overview.** In our self-supervised approach, we (a) pre-train an initial interest point detector on synthetic data and (b) apply a novel Homographic Adaptation procedure to automatically label images from a target, unlabeled domain. The generated labels are used to (c) train a fully-convolutional network that jointly extracts interest points and descriptors from an image.

[PDF] SuperPoint: Self-Supervised Interest Point Detection and Description D DeTone, T Malisiewicz, A Rabinovich - arXiv preprint arXiv:1712.07629, 2017

Keypoint-based method: Baseline



- Nearest neighbor
  - SIFT: Euclidean distance

Keypoint-based method: Baseline

Input | Detection | Description | Matching | **Estimation** | Output

- DSAC-differentiable RANSAC for camera localization

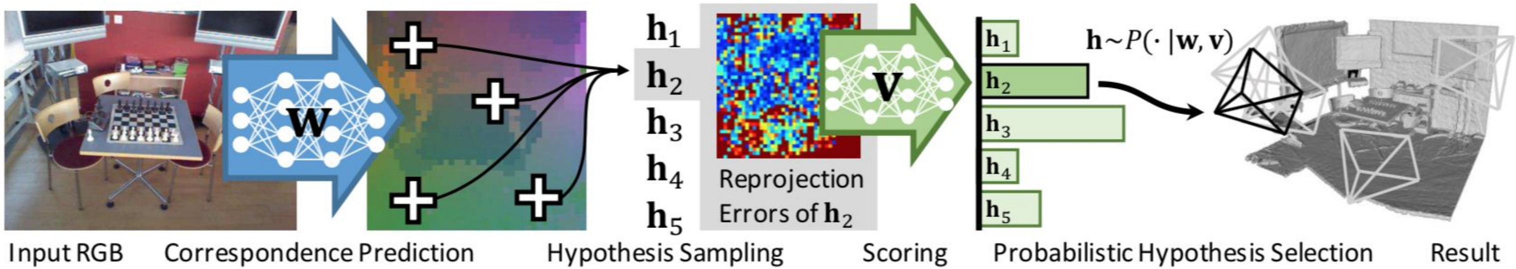


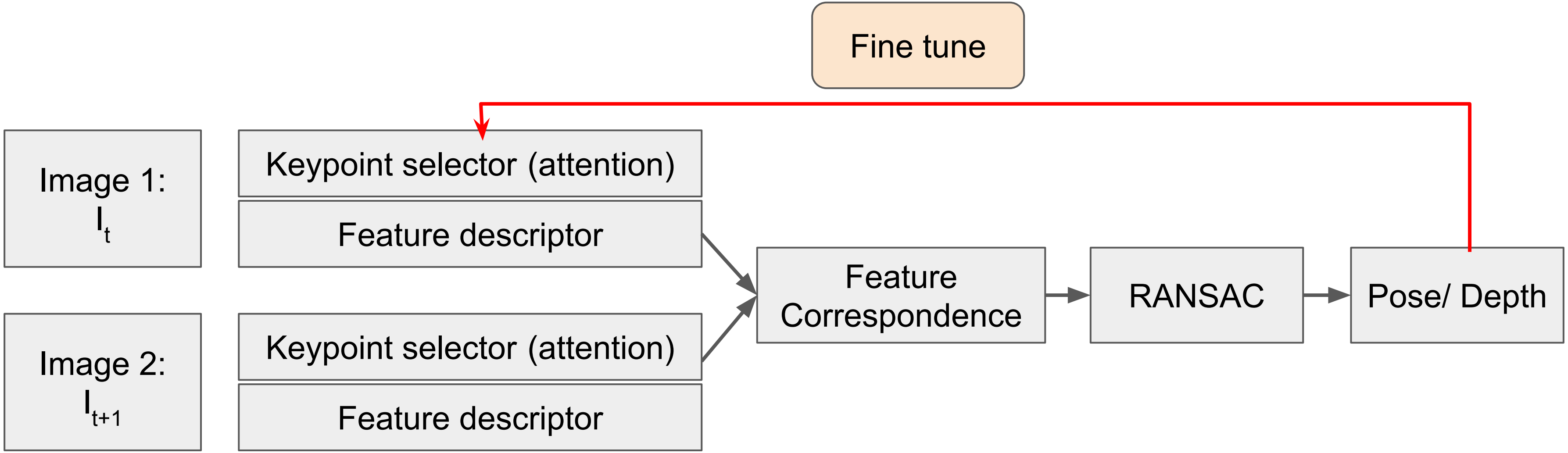
Figure 2. **Differentiable Camera Localization Pipeline.** Given an RGB image, we let a CNN with parameters $\mathbf{w}$ predict 2D-3D correspondences, so called scene coordinates [36]. From these, we sample minimal sets of four scene coordinates and create a pool of hypotheses $\mathbf{h}$. For each hypothesis, we create an image of reprojection errors which is scored by a second CNN with parameters $\mathbf{v}$. We select a hypothesis probabilistically according to the score distribution. The selected pose is also refined.

[PDF] DSAC-differentiable RANSAC for camera localization

E Brachmann, A Krull, S Nowozin, J Shotton, F Michel… - IEEE Conference on …, 2017

Keypoint-based method: Our method

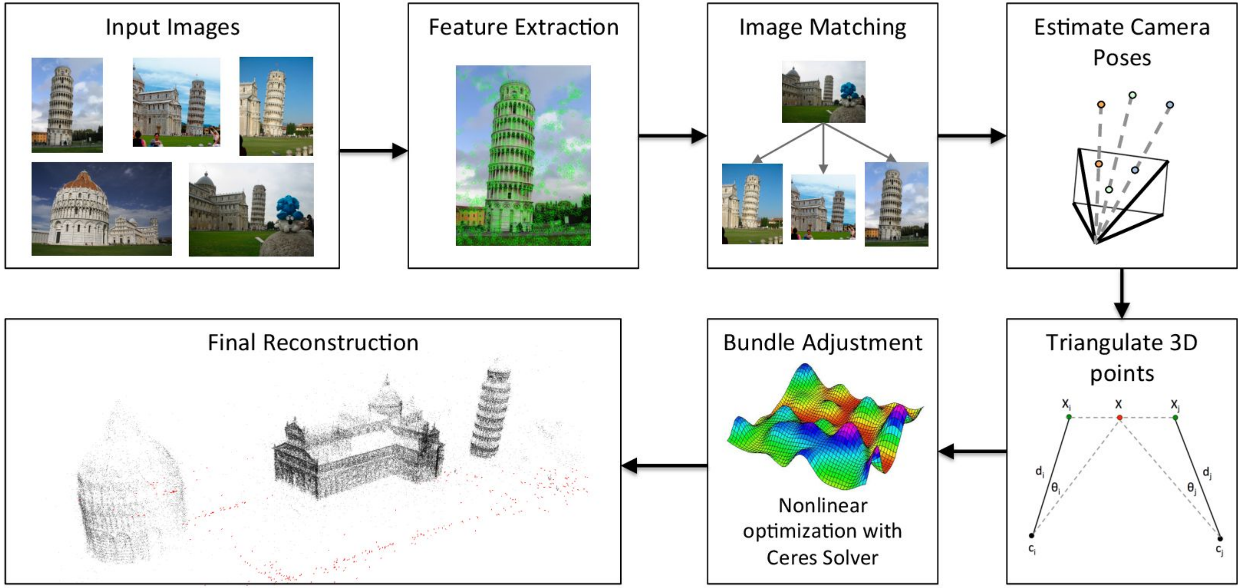
- Goal
  - Optimize the modules based on output error (keypoint reprojection error)
  - **Self-supervised** training using estimated pose and depth.
  - Use temporal information to **adapt keypoint selector** (attention map)



Backup slides

Reference Pipeline

http://www.theia-sfm.org/sfm.html



SuperPoint: Self-Supervised Interest Point Detection and Description

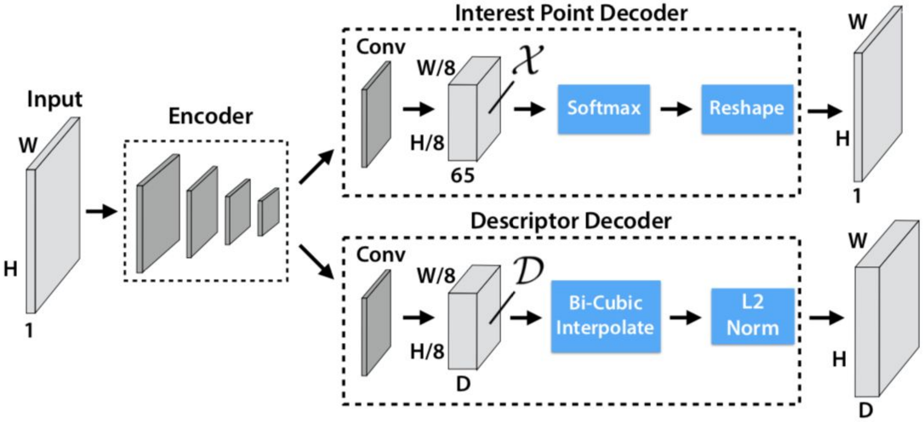


Figure 3. **SuperPoint Decoders**. Both decoders operate on a shared and spatially reduced representation of the input. To keep the model fast and easy to train, both decoders use non-learned upsampling to bring the representation back to $\mathbb{R}^{H \times W}$.

Comparison between relevant methods

	Interest Points?	Descriptors?	Full Image Input?	Single Network?	Real Time?
SuperPoint (ours)	✓	✓	✓	✓	✓
LIFT [32]	✓	✓			
UCN [3]		✓	✓	✓	
TILDE [29]	✓			✓	
DeepDesc [6]		✓		✓	
SIFT	✓	✓			
ORB	✓	✓			✓

Table 1. **Qualitative Comparison to Relevant Methods.** Our SuperPoint method is the only one to compute both interest points and descriptors in a single network in real-time.