Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?

Torsten Sattler[1] Akihiko Torii[2] Josef Sivic[3,5] Marc Pollefeys[1,4] Hajime Taira[2] Masatoshi Okutomi[2] Tomas Pajdla[5]

Contributions

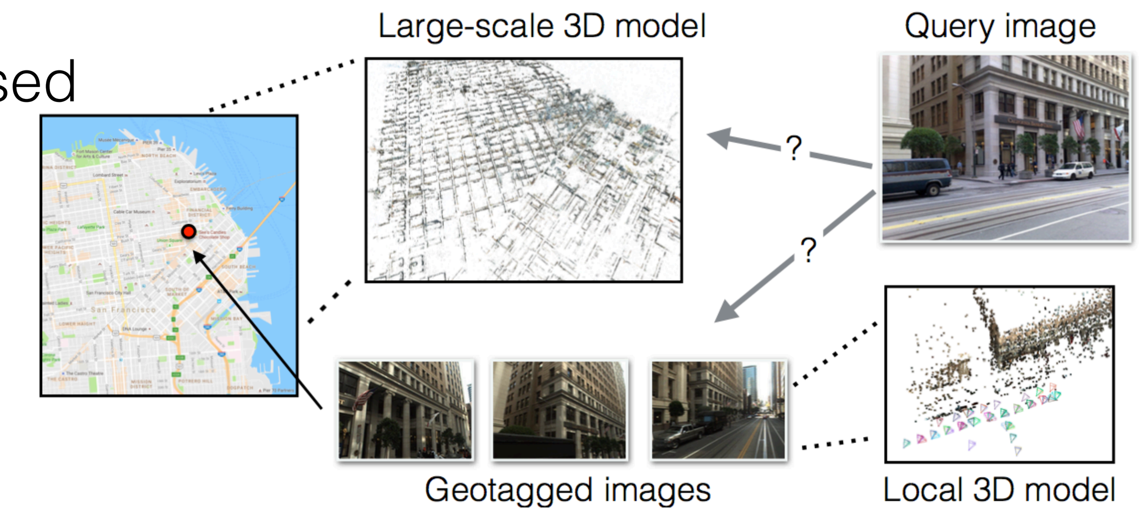
1. First **reference camera pose dataset** for large-scale localization (annotations for the query images of the San Francisco dataset [3])
2. First **comparison of 2D- and 3D**-localization **approaches** regarding their pose accuracy
3. Insight that **accurate visual localization** is possible **without large-scale 3D models** via **2D image retrieval** and **local SfM**

Accurate Visual Localization

Motivation: Self-driving cars, robots, AR

Two major approaches:

- **Accurate**: 3D structure-based localization via SfM models
- **Approximate**: 2D image retrieval-based localization

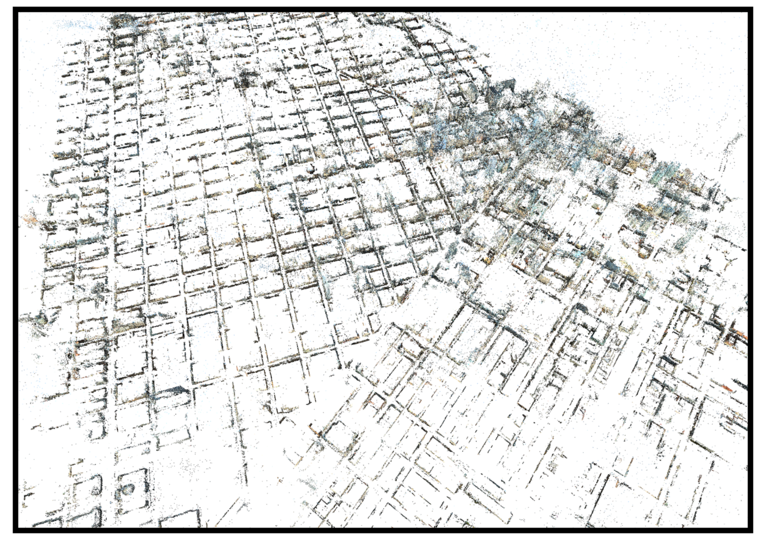


Challenges:

1. No large-scale dataset with ground truth poses
   - **San Francisco Landmarks** dataset [3]:

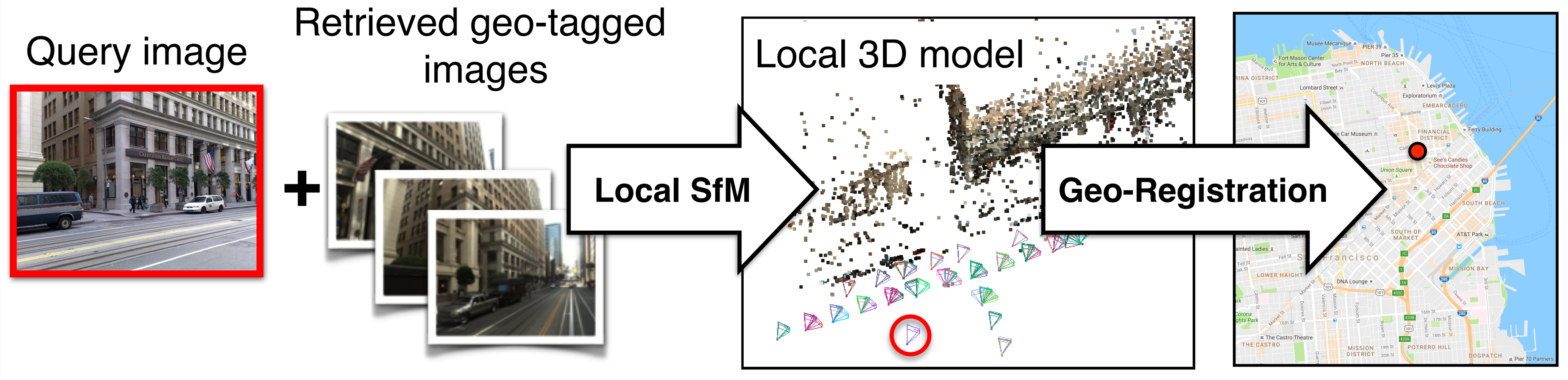
Database (DB): 1,06M PCI images, per image: building ID, accurate GPS	Query: 803 mobile phone photos, per query: building IDs, very inaccurate GPS

2. **Constructing**, **maintaining**, and storing **large SfM model**:
   - Adding or removing images requires **refinement via Bundle Adjustment**
   - **SF-0** SfM model [5]: 611k images, 30M 3D points



Are Large-Scale 3D Models Really Necessary?

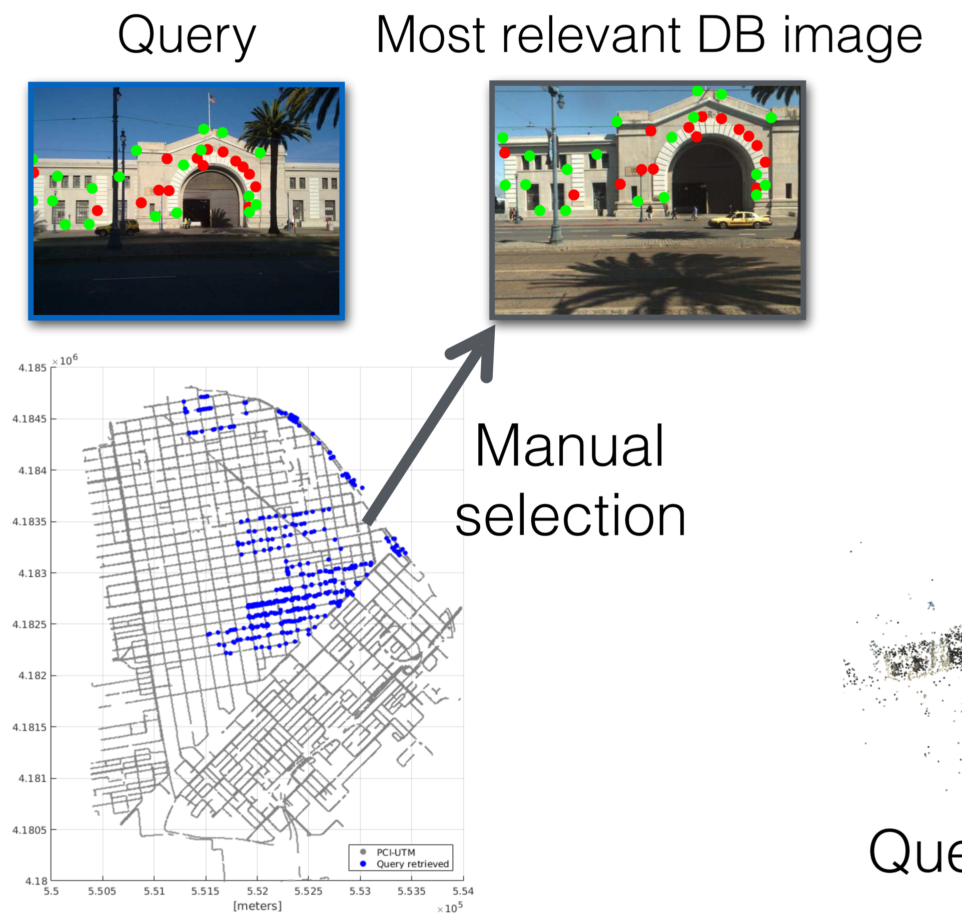
- **Constructing** & **maintaining image database** very easy
- Small memory footprint via compact image descriptors (≤16KB per image)
- **Approximate pose**: **Image retrieval** and known poses of database images
- **Accurate pose** estimation via post-processing: **Local SfM**+geo-registration



Reference Poses for San Francisco

http://www.ok.sc.e.titech.ac.jp/~torii/project/vlocalization/

Results | **Reference Poses** | **Benchmark Protocol**

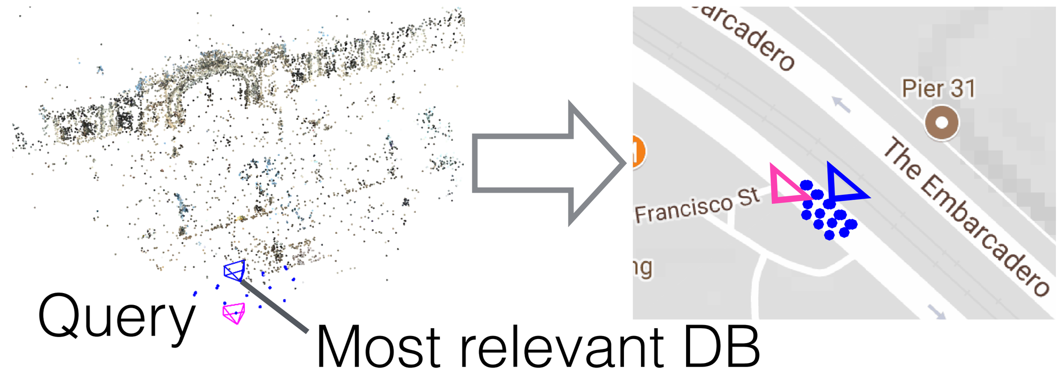


Manual annotation:

a. 2D-2D correspondences (green)
b. 2D-3D correspondences (red)

Pose estimation:

1. Local SfM (query + db. images)
2. Geo-registration using GPS of DB



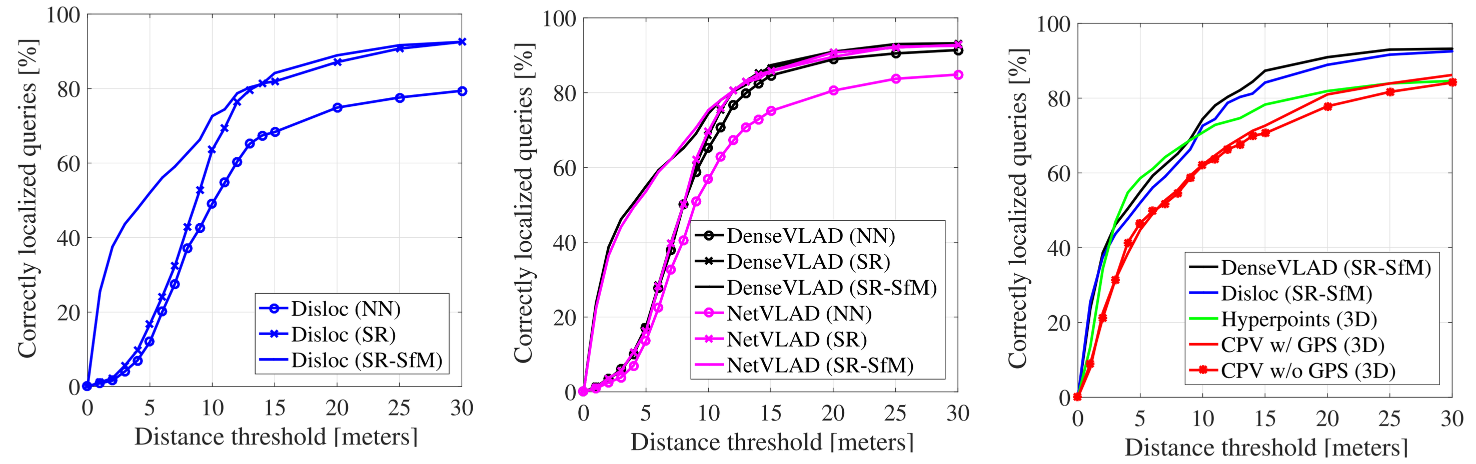
"Reference poses" consistent with manual annotations

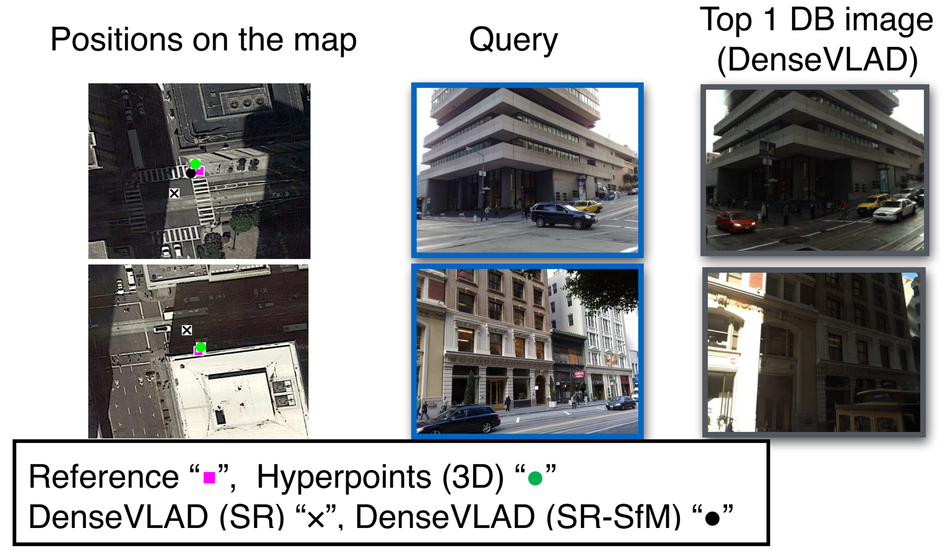
Main Insights

- **Large-scale 3D models not necessary** for accurate visual localization
- Accurate localization possible by combining **image retrieval + local SfM**, at the **price of run-time**
- **Retrieval** can **succeed** where **pose estimation fails** due to lack of matches
- Global 3D models can provide more accurate estimates for **some cases** where **local SfM is inaccurate / unstable** → research on **robust SfM**

Experiments

- 2D retrieval-based: NetVLAD [1], Disloc [2] + geom. burstiness [7], DenseVLAD [9]
- 3D-based: Hyperpoints [6], Active Search [8], Camera Pose Voting (CPV) [10]
- Variants for 2D: Nearest neighbor (**NN**), spatial verification (**SR**), local SfM (**SfM**)
- **Evaluation measure**: Percentage of query images with pose **within X meters** of reference pose, distances measured in UTM coordinates **in 2D** (height undefined)
- Results for all **San Francisco references** poses:





- Results for **Dubrovnik** dataset (6044 db. images, 1.9M 3D points):

Method	Time [sec]	Quantile errors [m] 25%	50%	75%
DenseVLAD [9] (NN)	1.42	1.4	3.9	11.2
DenseVLAD [9] (SR)	1.43	0.9	2.9	9.0
DenseVLAD [9] (SR-SfM)	~200	0.3	1.0	5.1
Camera Pose Voting (CPV) [10]	3.78	0.19	0.56	2.09
Active Search [8]	0.16	0.5	1.3	5.0
PoseNet [4]	~0.005	-	7.9	-

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. CVPR, 2016.
[2] R. Arandjelović and A. Zisserman. DisLocation: Scalable descriptor distinctiveness for location recognition. ACCV, 2014.
[3] D. Chen, G. Baatz, K. Köser, et. al. City-scale landmark identification on mobile devices. CVPR, 2011.
[4] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. CVPR, 2017.
[5] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide Pose Estimation Using 3D Point Clouds. ECCV, 2012
[6] T. Sattler, M. Havlena, F. Radenović, et. al. Hyperpoints and fine vocabularies for large-scale location recognition. ICCV, 2015.
[7] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-Scale Location Recognition and the Geometric Burstiness Problem. CVPR, 2016.
[8] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. PAMI, 2016
[9] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. CVPR, 2015.
[10] B. Zeisl, T. Sattler, and M. Pollefeys. Camera pose voting for large-scale image-based localization. ICCV, 2015.

[1]ETH Zürich [2]Tokyo Institute of Technology [3]Inria, DI ENS, PSL Research University [4]Microsoft Corp. [5]CIIRC, Czech Technical University in Prague

This work was partly supported by EU-H2020 project LADIO No. 731970, JSPS KAKENHI Grant Number 15H05313, ERC grant LEAP (no. 336845), CIFAR Learning in Machines & Brains program and ESIF, OP Research, development and education project IMPACT No. CZ.02.1.01/0.0/0.0/15 003/0000468, and Google Tango.