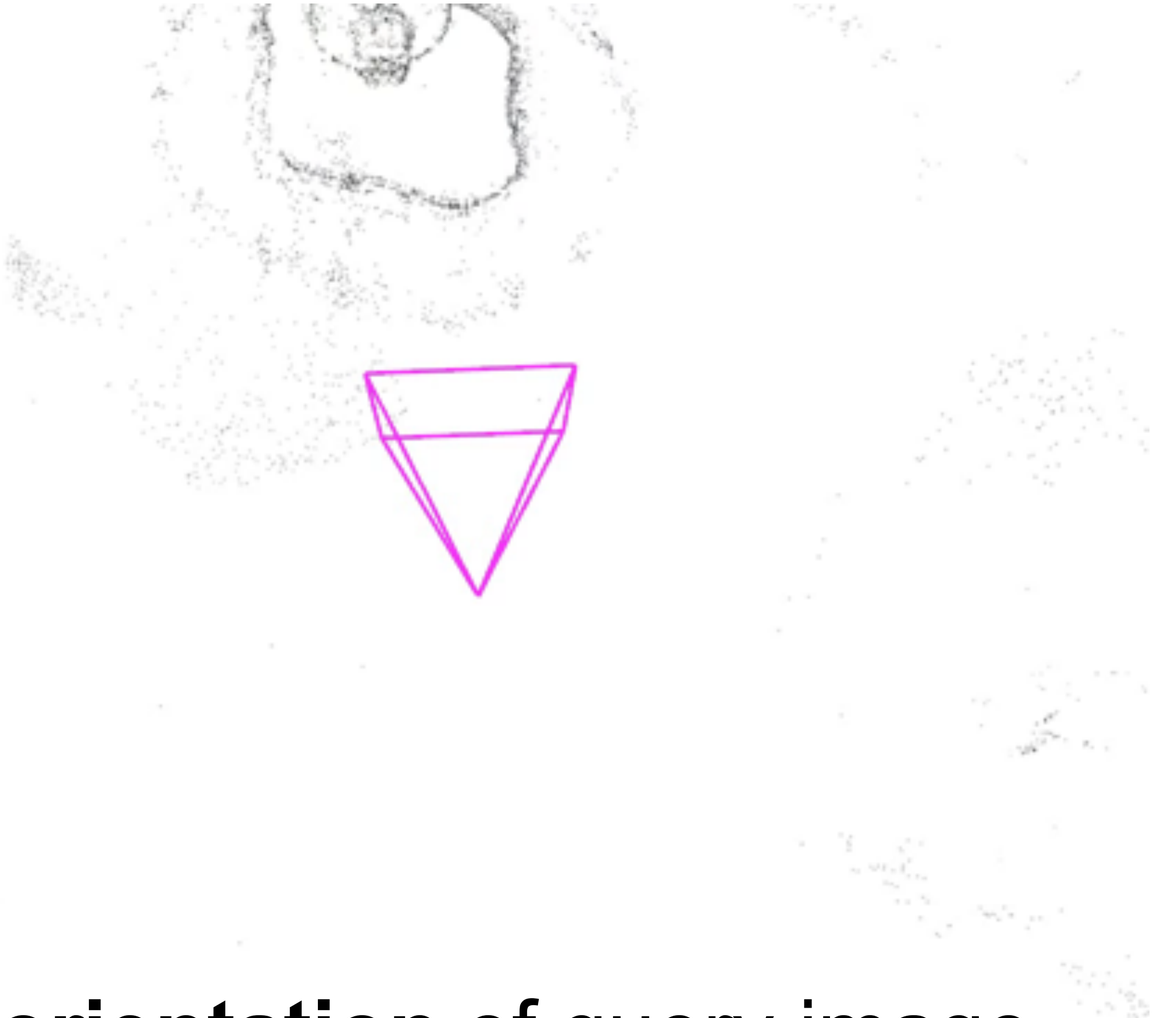
In Defense of Local Features for Visual Localization

Torsten Sattler

Department of Electrical Engineering
Chalmers University of Technology

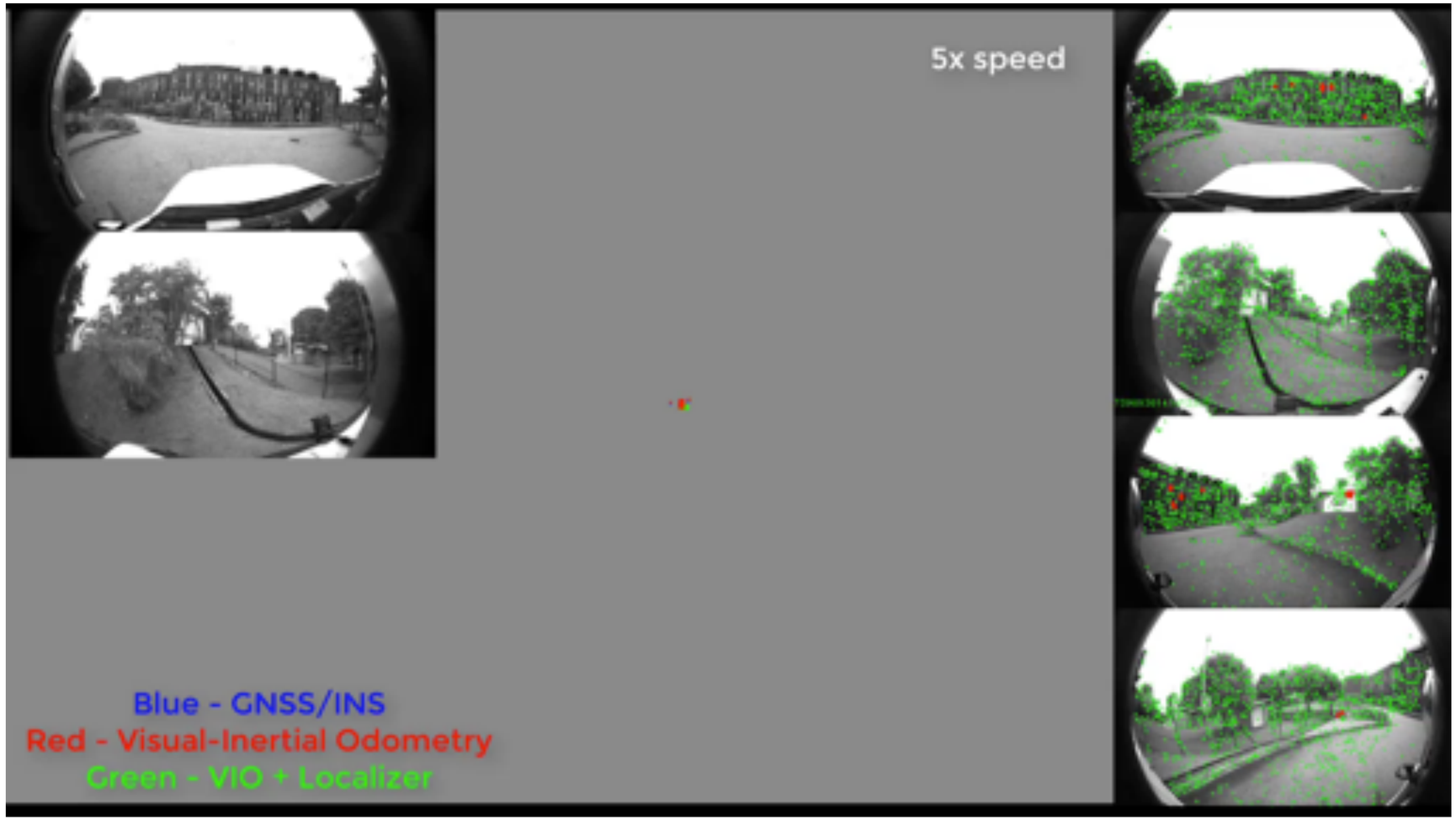
(work done at ETH Zürich)

The Visual Localization Problem



Compute **exact position and orientation** of query image

Applications: Autonomous Driving



AutoVision
3D Vision for Autonomous Vehicles

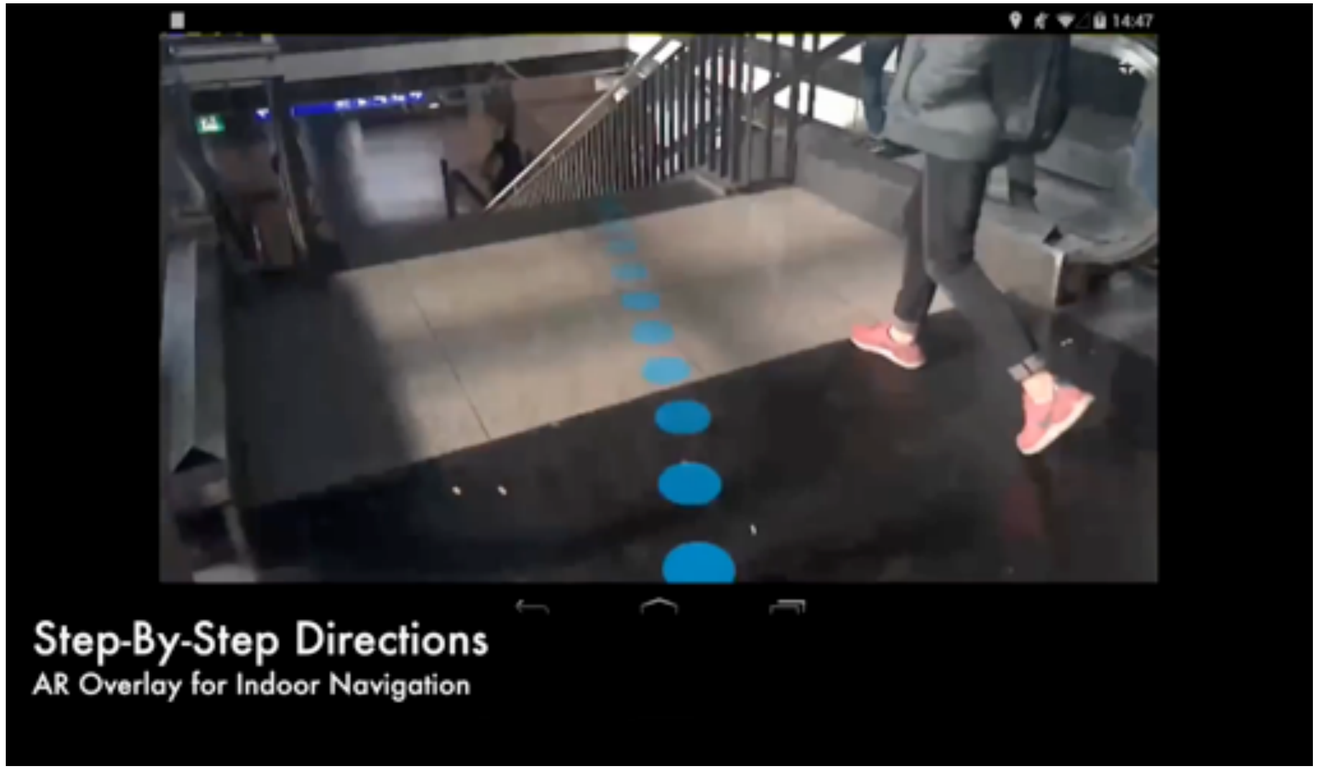
[Geppert, Liu, Cui, Pollefeys, Sattler, Efficient 2D-3D Matching for Multi-Camera Visual Localization, ICRA 2019]

Applications: Robotics

Horizon 2020
European Union funding
for Research & Innovation

Applications: Augmented Reality

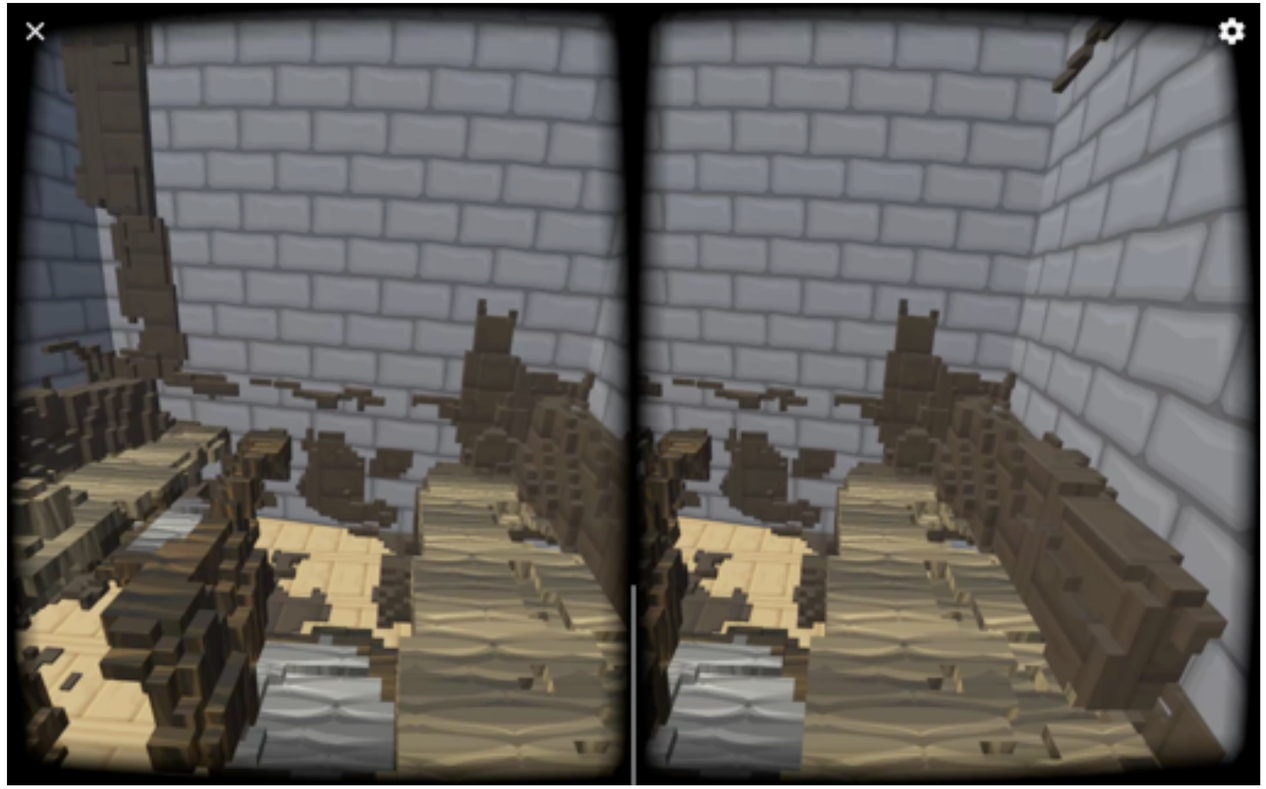
[Middelberg, Sattler, Untzelmann, Kobbelt, Scalable 6-DOF Localization on Mobile Devices, ECCV 2014]



Tango

[Google Tango video], © Google

Applications: Virtual Reality



Master thesis by Benjamin Steger, ETH Zurich
Textures: PureBDcraft ResourcePack by Sphax - BDcraft.net

Visual Localization Approaches

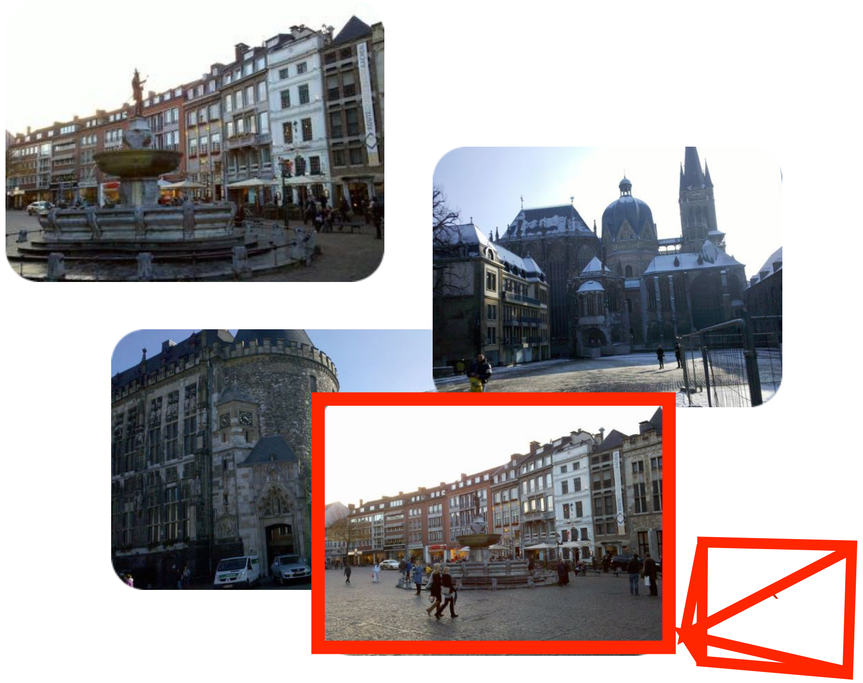
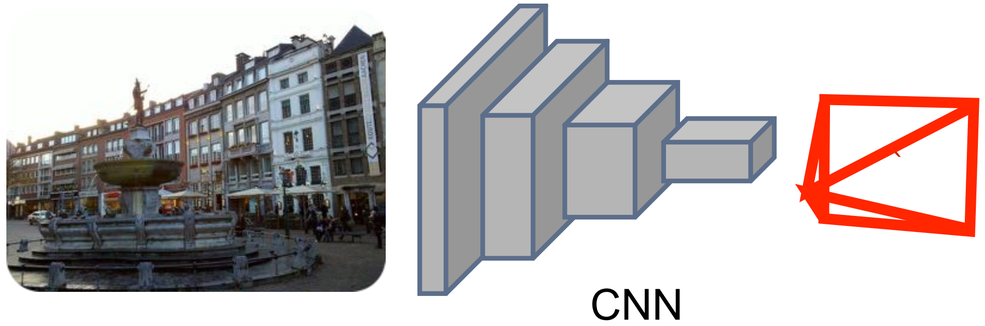


Image Retrieval

Structure-based



Pose Regression

Visual Localization Approaches

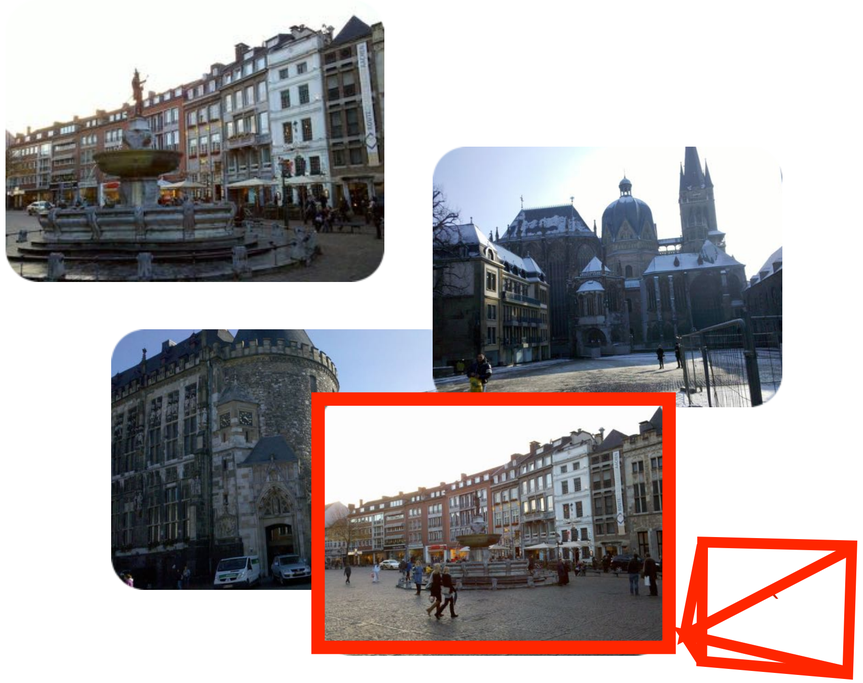
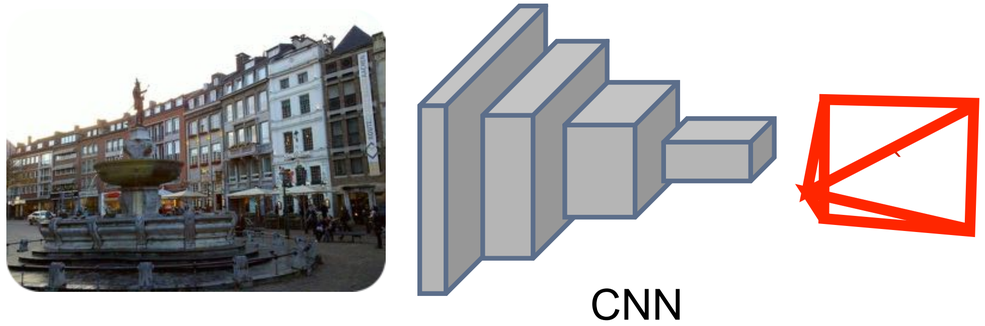


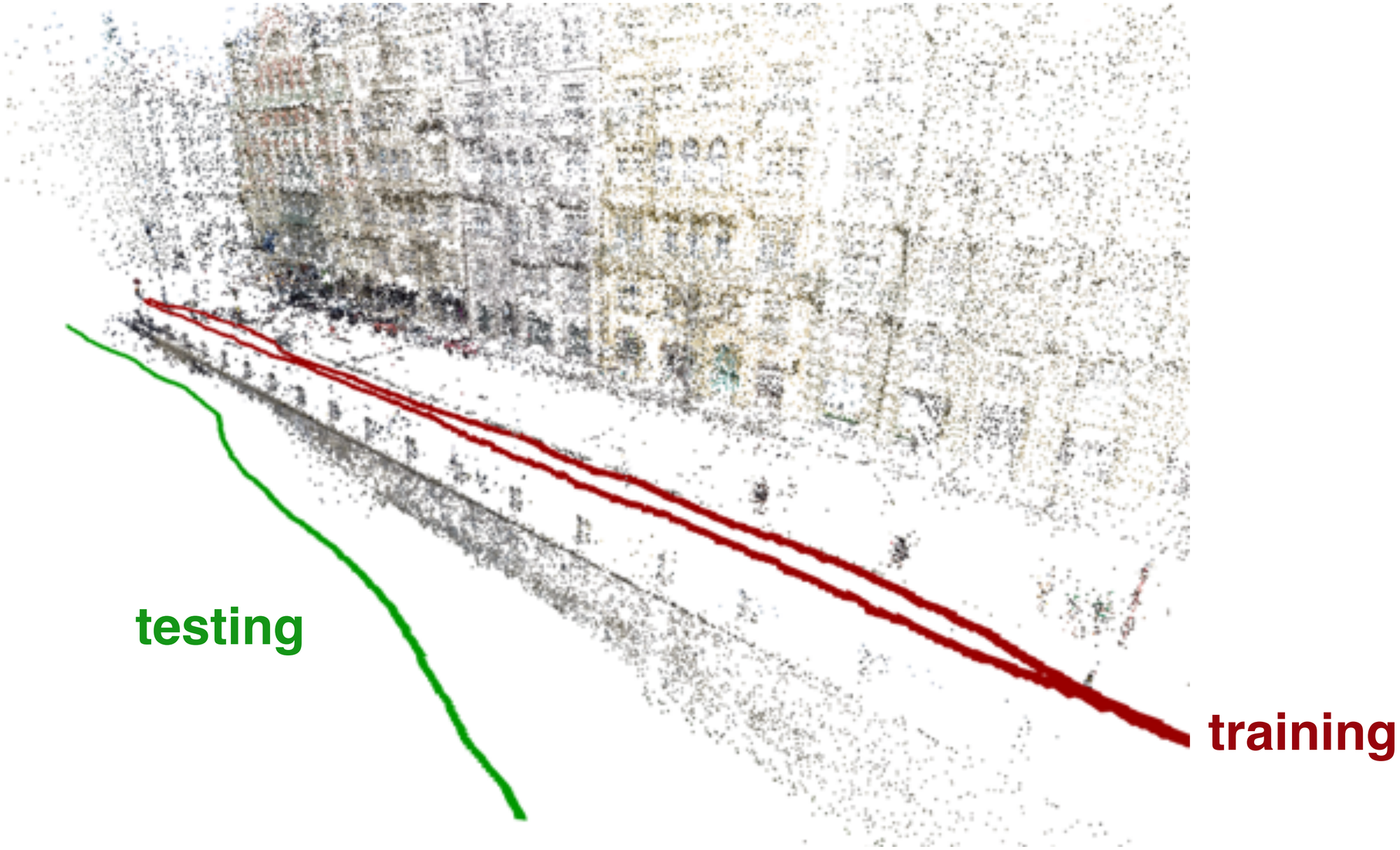
Image Retrieval

Structure-based



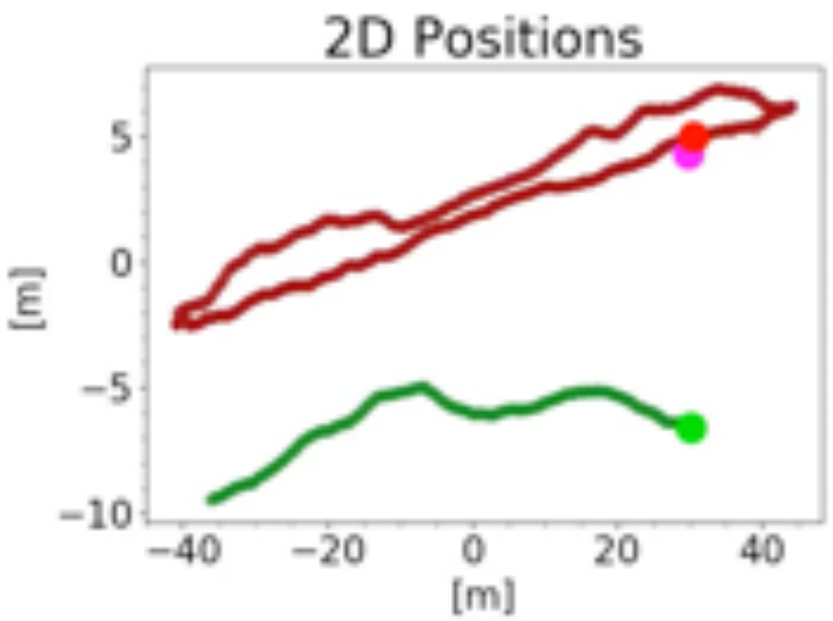
Pose Regression

Camera Pose Regression Example



[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Camera Pose Regression Example



GT training poses

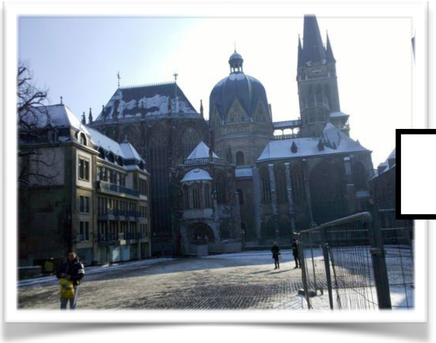
GT test poses

Predicted test pose

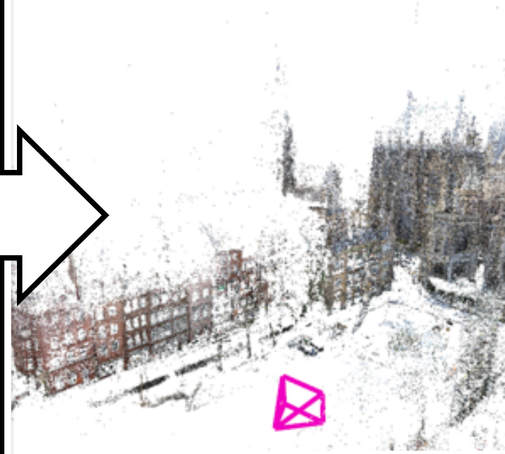
Pose of most similar training image

[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Looking Inside The Black Box

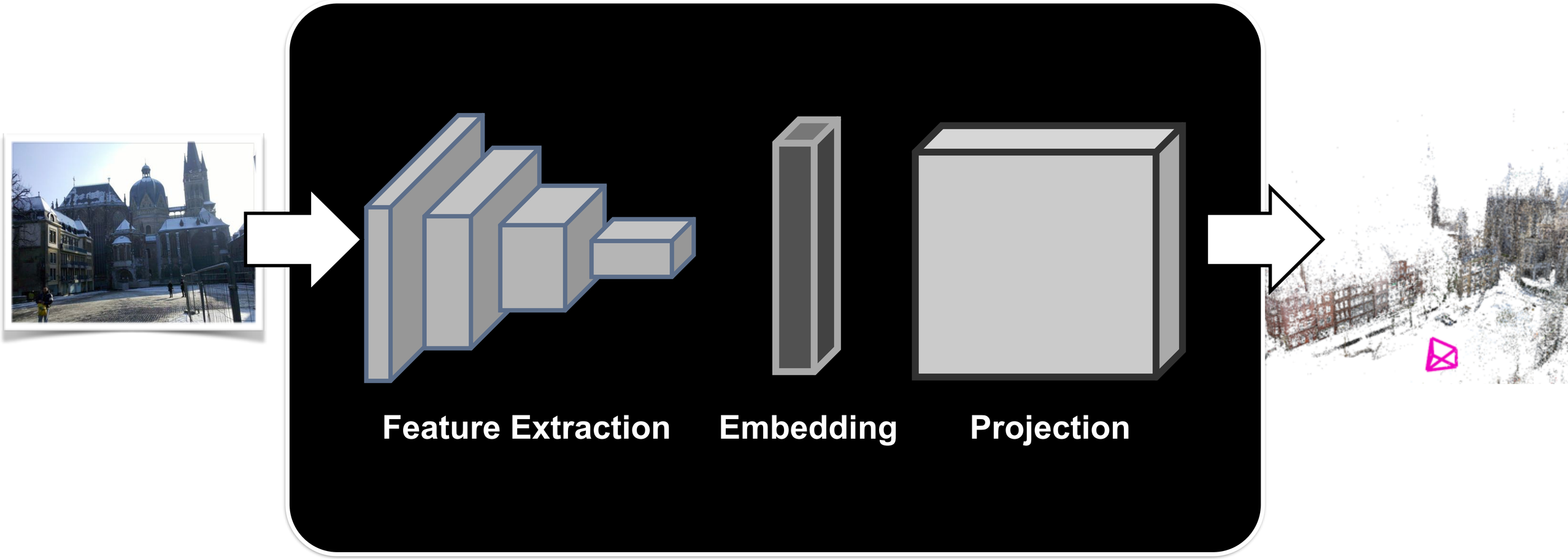


Convolutional Neural Network



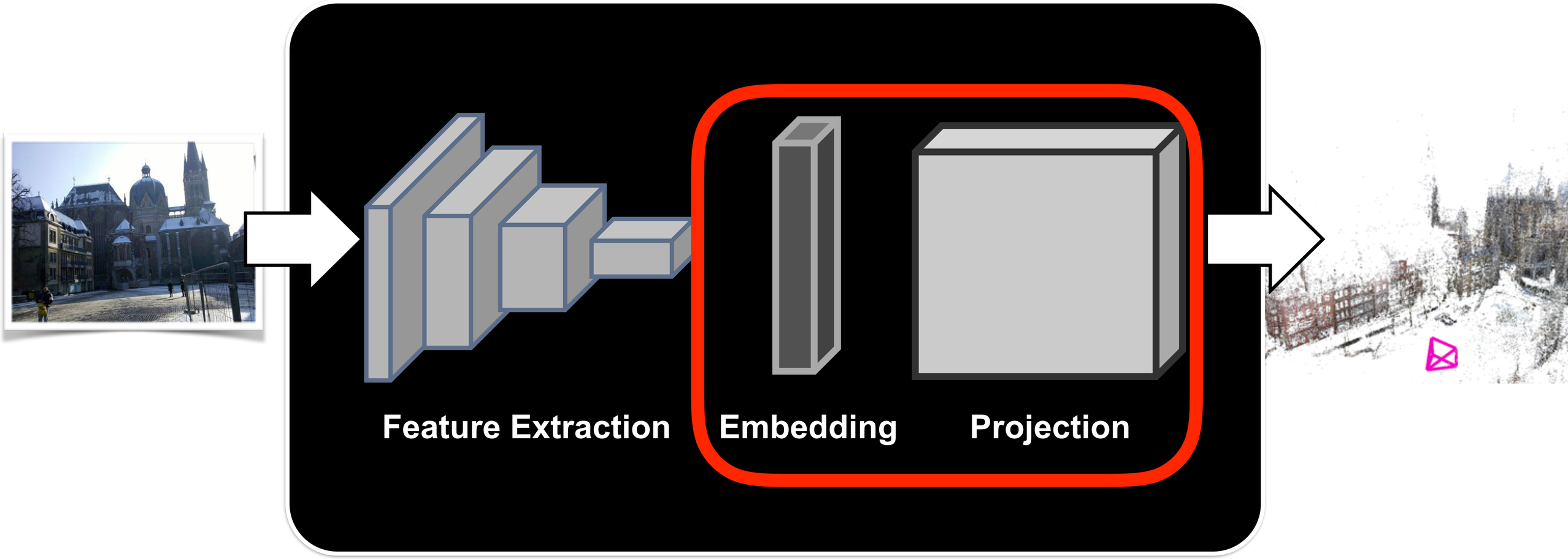
[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Looking Inside The Black Box



[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

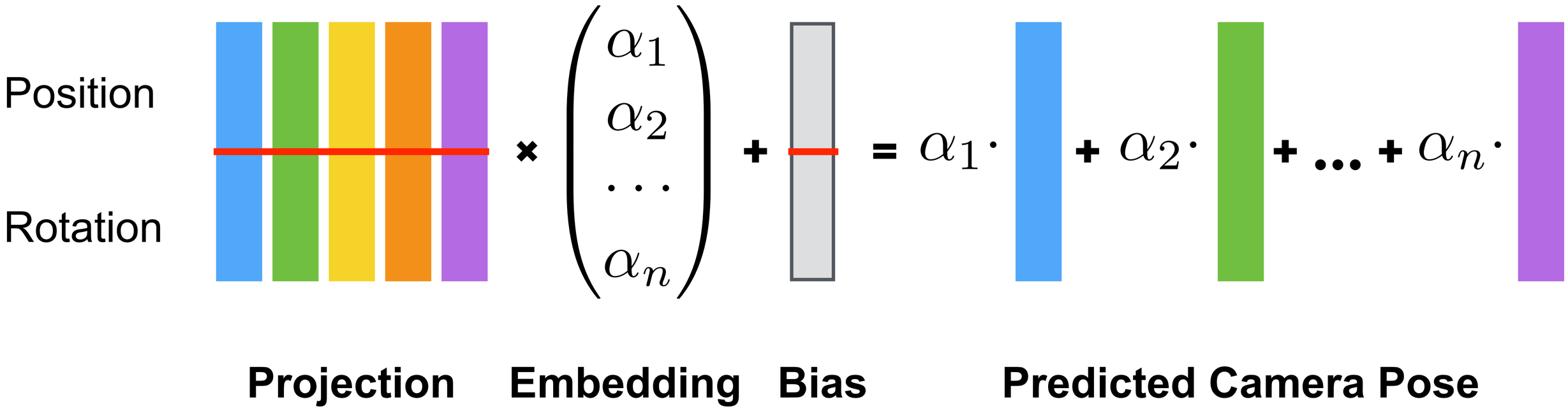
Looking Inside The Black Box



[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

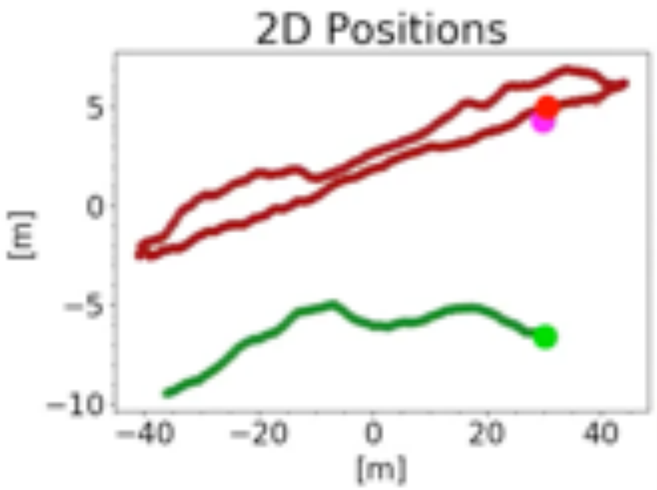
Looking Inside The Black Box

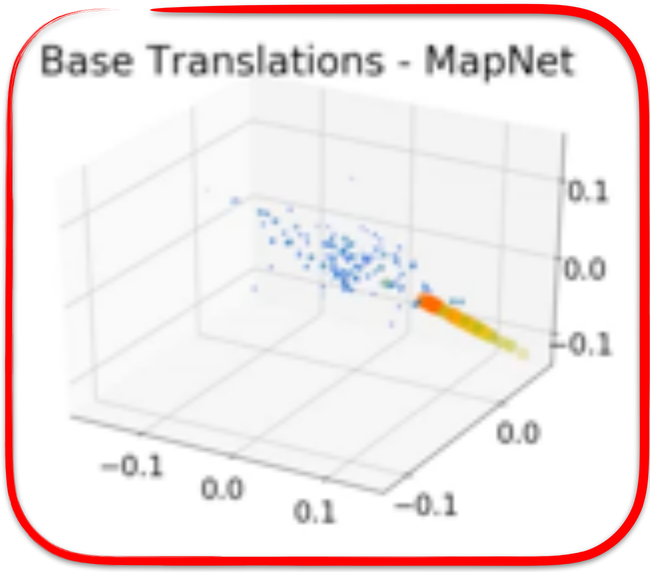
- Pose regression in **last FC layer** as **linear combination of base poses**:



[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Camera Pose Regression Example





GT training poses

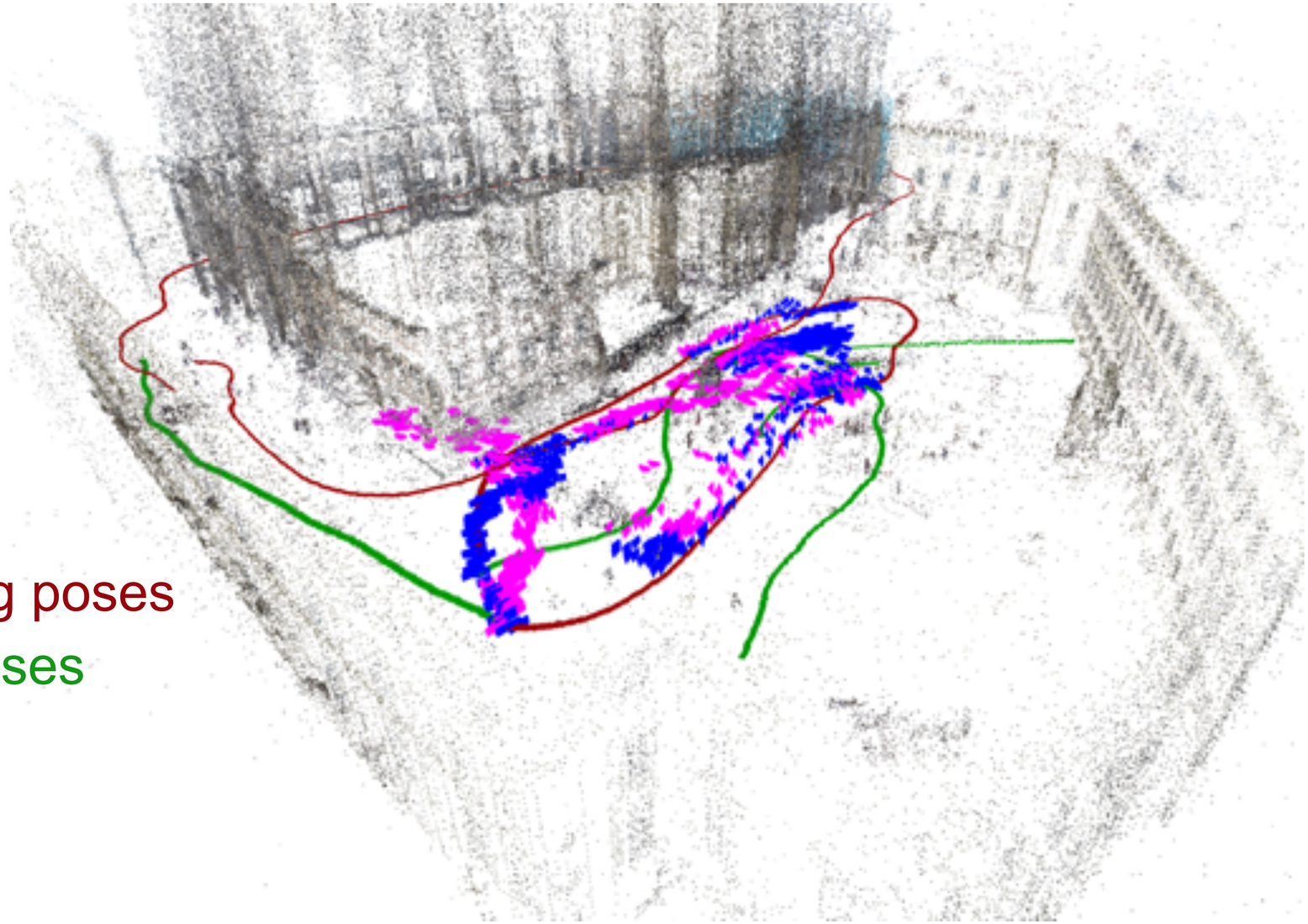
GT test poses

Predicted test pose

Pose of most similar training image

[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Camera Pose Regression Example

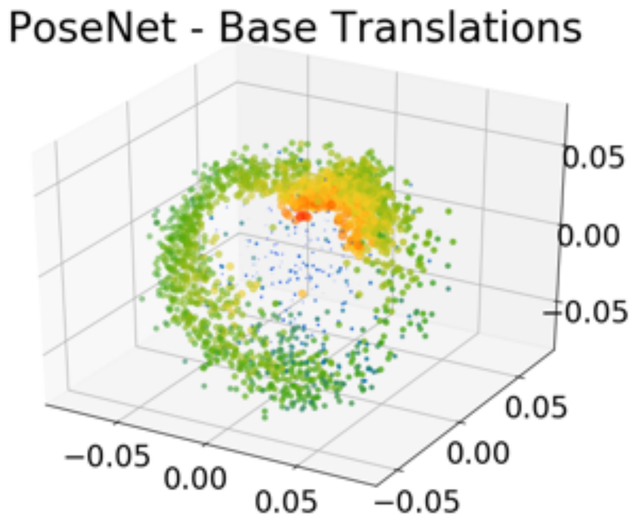


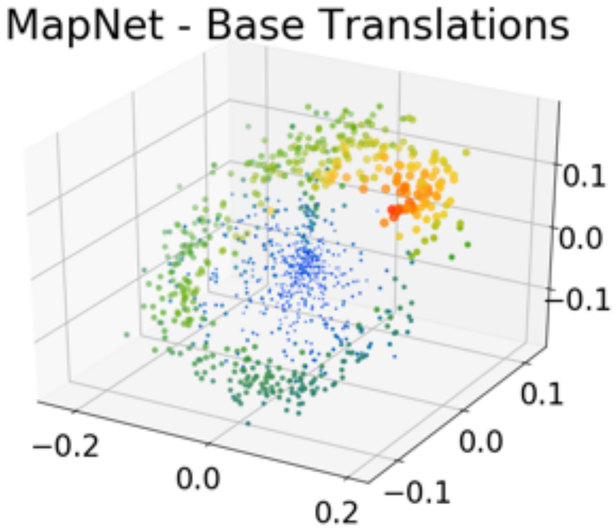
GT training poses

GT test poses

MapNet

PoseNet





[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Comparison to Image Retrieval

	Cambridge Landmarks					7 Scenes						
	Kings	Old	Shop	St. Mary's	Street	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
Camera Pose Regression												
Image Retrieval												
Structure-based Methods												

[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Comparison to Image Retrieval

		Cambridge Landmarks					7 Scenes						
		Kings	Old	Shop	St. Mary's	Street	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
Camera Pose Regression													
Image Retrieval													
3D	Active Search [59]	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54	**0.85/0.8**	0.04/1.96	0.03/1.53	0.02/1.45	0.09/3.61	0.08/3.10	0.07/3.37	**0.03/2.22**
	BTBRF [46]	0.39/0.36	0.30/0.41	0.15/0.31	0.20/**0.40**								
	DSAC++ [10]	**0.18/0.3**	**0.20/0.3**	**0.06/0.3**	**0.13/0.4**		**0.02/0.5**	**0.02/0.9**	**0.01/0.8**	**0.03/0.7**	**0.04/1.1**	**0.04/1.1**	0.09/2.6
	InLoc [69]						0.03/1.05	0.03/1.07	0.02/1.16	**0.03**/1.05	0.05/1.55	**0.04**/1.31	0.09/2.47

[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Comparison to Image Retrieval

Camera Pose Regression

		Cambridge Landmarks					7 Scenes						
		Kings	Old	Shop	St. Mary's	Street	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
IR	DenseVLAD [71]	2.80/5.72	4.01/7.13	1.11/7.61	2.31/8.00	5.16/23.5	0.21/12.5	0.33/13.8	0.15/14.9	0.28/11.2	0.31/11.3	0.30/12.3	0.25/15.8
IR	DenseVLAD + Inter.	1.48/4.45	2.68/4.63	0.90/4.32	1.62/6.06	15.4/25.7	0.18/10.0	0.33/12.4	0.14/14.3	0.25/10.1	0.26/9.42	0.27/11.1	0.24/14.7
3D	Active Search [59]	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54	**0.85/0.8**	0.04/1.96	0.03/1.53	0.02/1.45	0.09/3.61	0.08/3.10	0.07/3.37	**0.03/2.22**
3D	BTBRF [46]	0.39/0.36	0.30/0.41	0.15/0.31	0.20/**0.40**								
3D	DSAC++ [10]	**0.18/0.3**	**0.20/0.3**	**0.06/0.3**	**0.13/0.4**		**0.02/0.5**	**0.02/0.9**	**0.01/0.8**	**0.03/0.7**	**0.04/1.1**	**0.04/1.1**	0.09/2.6
3D	InLoc [69]						0.03/1.05	0.03/1.07	0.02/1.16	**0.03**/1.05	0.05/1.55	**0.04**/1.31	0.09/2.47

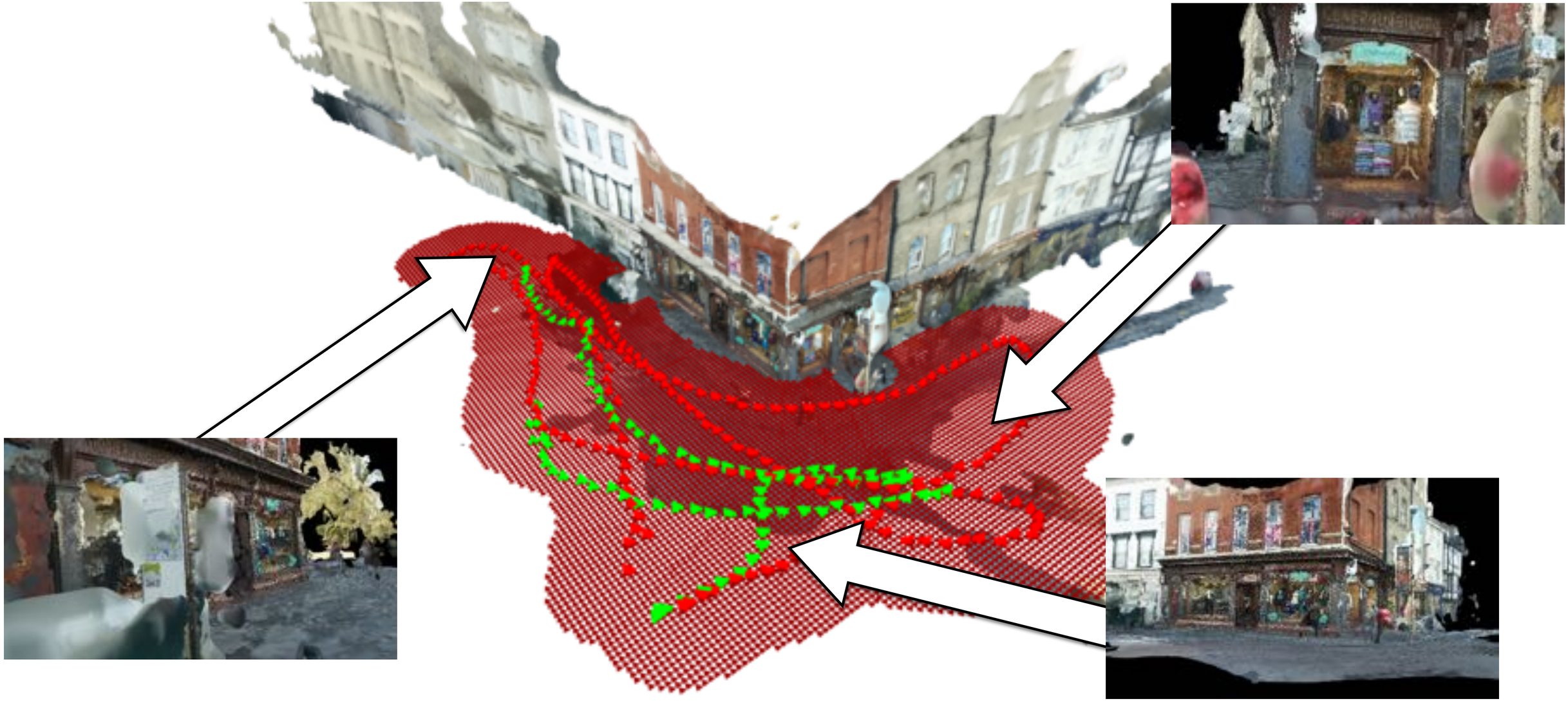
[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Comparison to Image Retrieval

		Cambridge Landmarks					7 Scenes						
		Kings	Old	Shop	St. Mary's	Street	Chess	Fire	Heads	Office	Pumpkin	Kitchen	Stairs
APR	PoseNet (PN) [30]	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48		0.32/8.12	0.47/14.4	0.29/12.0	0.48/7.68	0.47/8.42	0.59/8.64	0.47/13.8
	PN learned weights [29]	0.99/1.06	2.17/2.94	1.05/3.97	1.49/3.43	20.7/25.7	0.14/4.50	0.27/11.8	0.18/12.1	0.20/5.77	0.25/4.82	0.24/5.52	0.37/10.6
	Bay. PN [28]	1.74/4.06	2.57/5.14	1.25/7.54	2.11/8.38		0.37/7.24	0.43/13.7	0.31/12.0	0.48/8.04	0.61/7.08	0.58/7.54	0.48/13.1
	geo. PN [29]	0.88/1.04	3.20/3.29	0.88/3.78	1.57/3.32	20.3/25.5	0.13/4.48	0.27/11.3	0.17/13.0	0.19/5.55	0.26/4.75	0.23/5.35	0.35/12.4
	LSTM PN [76]	0.99/3.65	1.51/4.29	1.18/7.44	1.52/6.68		0.24/5.77	0.34/11.9	0.21/13.7	0.30/8.08	0.33/7.00	0.37/8.83	0.40/13.7
	GPoseNet [12]	1.61/2.29	2.62/3.89	1.14/5.73	2.93/6.46		0.20/7.11	0.38/12.3	0.21/13.8	0.28/ 8.83	0.37/6.94	0.35/8.15	0.37/12.5
	SVS-Pose [50]	1.06/2.81	1.50/4.03	0.63/5.73	2.11/8.11								
	Hourglass PN [44]						0.15/6.17	0.27/10.8	0.19/11.6	0.21/8.48	0.25/7.01	0.27/10.2	0.29/12.5
	BranchNet [78]						0.18/5.17	0.34/8.99	0.20/14.2	0.30/7.05	0.27/5.10	0.33/7.40	0.38/10.3
	MapNet [11]	1.07/1.89	1.94/3.91	1.49/4.22	2.00/4.53		0.08/3.25	0.27/11.7	0.18/13.3	0.17/5.15	0.22/4.02	0.23/4.93	0.30/12.1
	MapNet+ [11]						0.10/3.17	0.20/9.04	0.13/11.1	0.18/5.38	0.19/3.92	0.20/5.01	0.30/13.4
	MapNet+PGO [11]						0.09/3.24	0.20/9.29	0.12/8.45	0.19/5.42	0.19/3.96	0.20/4.94	0.27/10.6
RPR	Relative PN [35] (U)						0.31/15.0	0.40/19.0	0.24/22.2	0.38/14.1	0.44/18.2	0.41/16.5	0.35/23.6
	Relative PN [35] (7S)						0.13/6.46	0.26/12.7	0.14/12.3	0.21/7.35	0.24/6.35	0.24/8.03	0.27/11.8
	RelocNet [7] (SN)						0.21/ 10.9	0.32/11.8	0.15/13.4	0.31/ 10.3	0.40/ 10.9	0.33/10.3	0.33/11.4
	RelocNet [7] (7S)						0.12/4.14	0.26/10.4	0.14/10.5	0.18/5.32	0.26/4.17	0.23/5.08	0.28/7.53
	AnchorNet [56]	0.57/0.88	1.21/2.55	0.52/2.27	1.04/2.69	7.86/24.2	0.06/3.89	0.15/10.3	0.08/10.9	0.09/5.15	0.10/2.97	0.08/4.68	0.10/9.26
IR	DenseVLAD [71]	2.80/5.72	4.01/7.13	1.11/7.61	2.31/8.00	5.16/23.5	0.21/12.5	0.33/13.8	0.15/14.9	0.28/11.2	0.31/11.3	0.30/12.3	0.25/15.8
	DenseVLAD + Inter.	1.48/4.45	2.68/4.63	0.90/4.32	1.62/6.06	15.4/25.7	0.18/10.0	0.33/12.4	0.14/14.3	0.25/10.1	0.26/9.42	0.27/11.1	0.24/14.7
3D	Active Search [59]	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54	**0.85/0.8**	0.04/1.96	0.03/1.53	0.02/1.45	0.09/3.61	0.08/3.10	0.07/3.37	**0.03/2.22**
	BTBRF [46]	0.39/0.36	0.30/0.41	0.15/0.31	0.20/**0.40**								
	DSAC++ [10]	**0.18/0.3**	**0.20/0.3**	**0.06/0.3**	**0.13/0.4**		**0.02/0.5**	**0.02/0.9**	**0.01/0.8**	**0.03/0.7**	**0.04/1.1**	**0.04/1.1**	0.09/2.6
	InLoc [69]						0.03/1.05	0.03/1.07	0.02/1.16	**0.03**/1.05	0.05/1.55	**0.04**/1.31	0.09/2.47

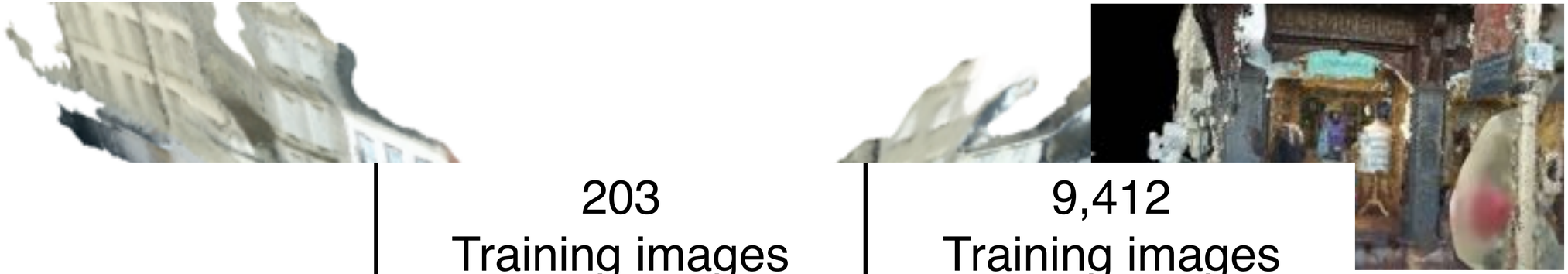
[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Training from More Data

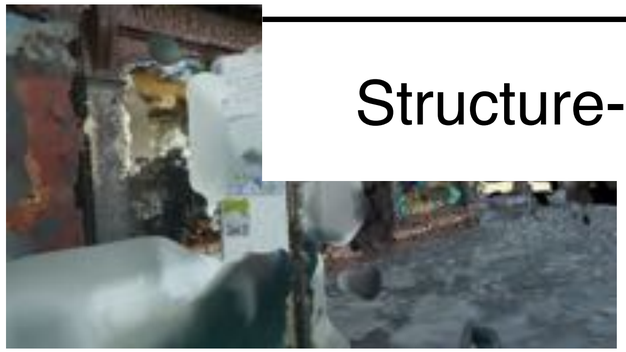


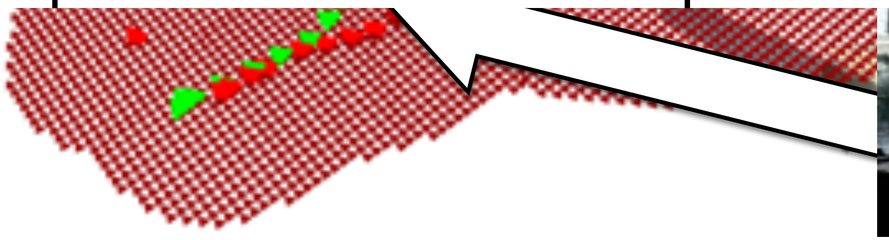
[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Training from More Data



	203 Training images	9,412 Training images
MapNet	1.07m / 4.70deg	0.33m / 1.46deg
Image Retrieval	0.89m / 5.71deg	0.38m / 6.41deg
Structure-Based	0.01m / 0.04deg	





[Sattler, Zhou, Pollefeys, Leal-Taixé, Understanding the Limitations of CNN-based Absolute Camera Pose Regression, CVPR 2019]

Visual Localization Approaches

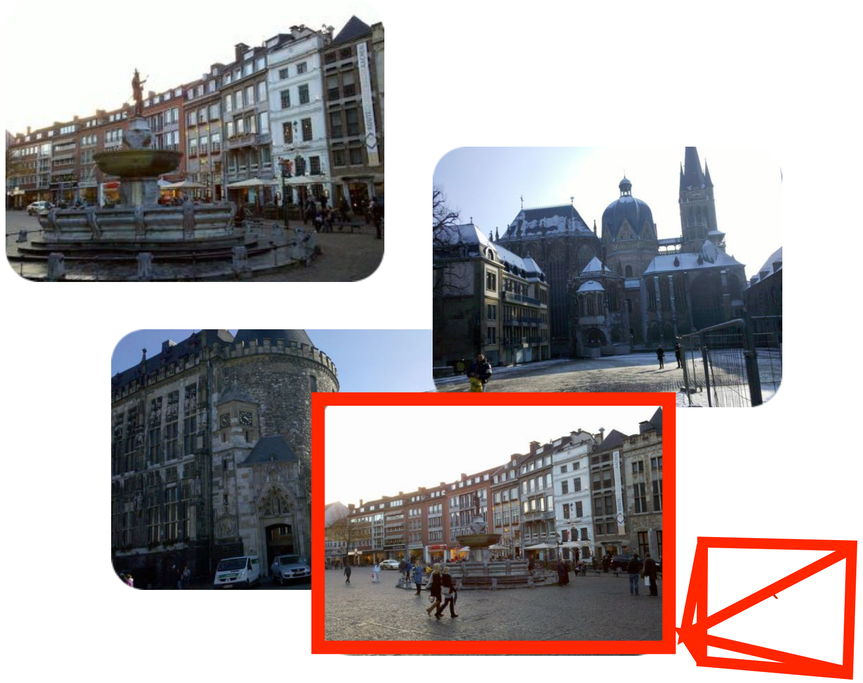
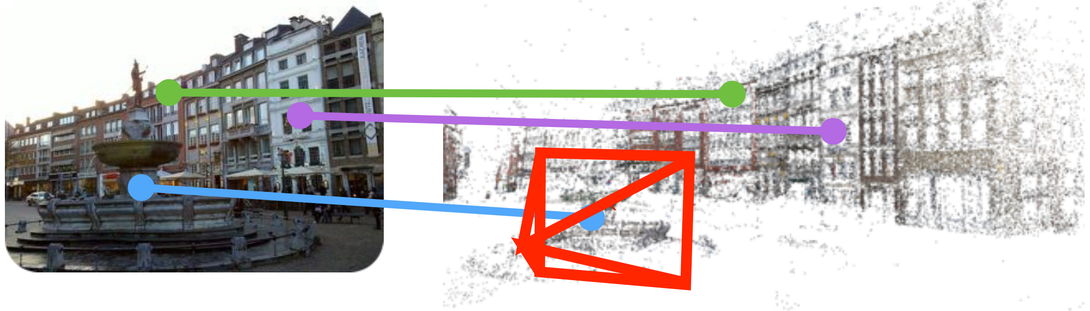


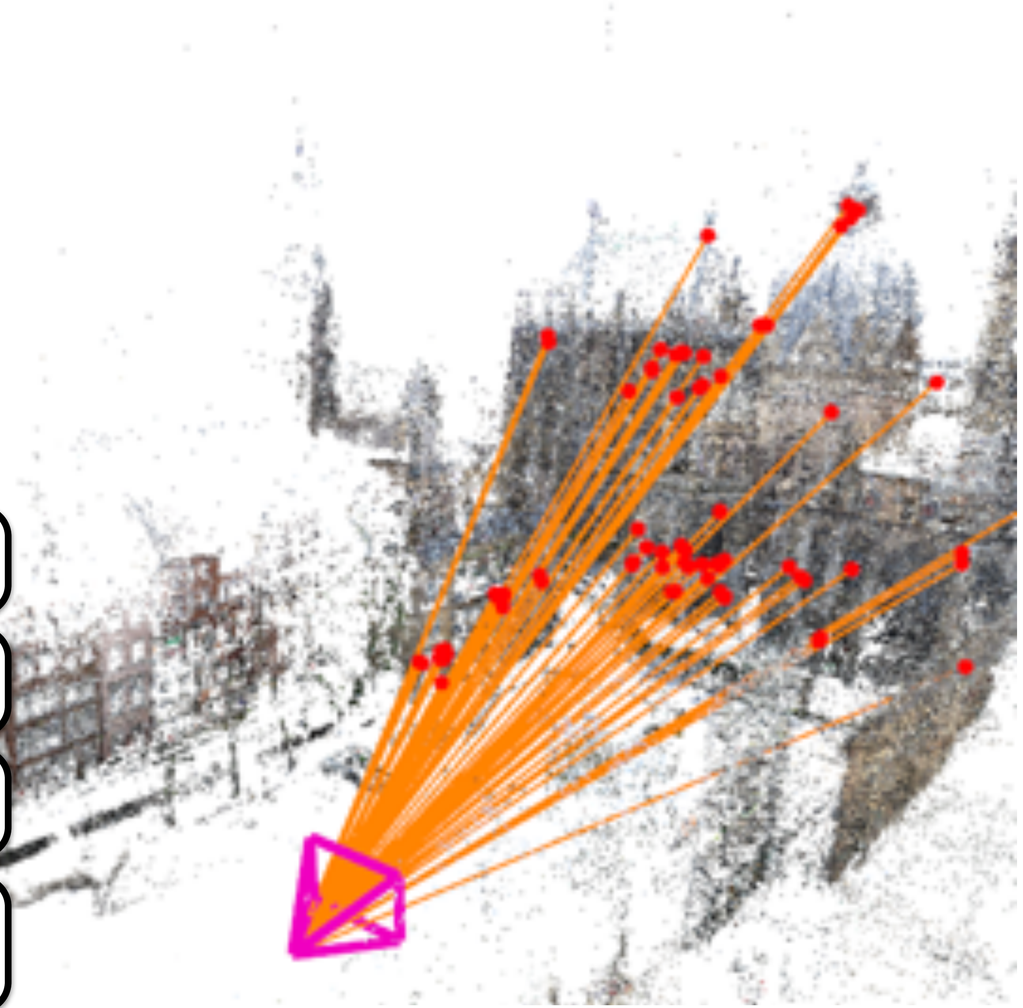
Image Retrieval



Structure-based

Pose Regression

Scene Coordinate Regression



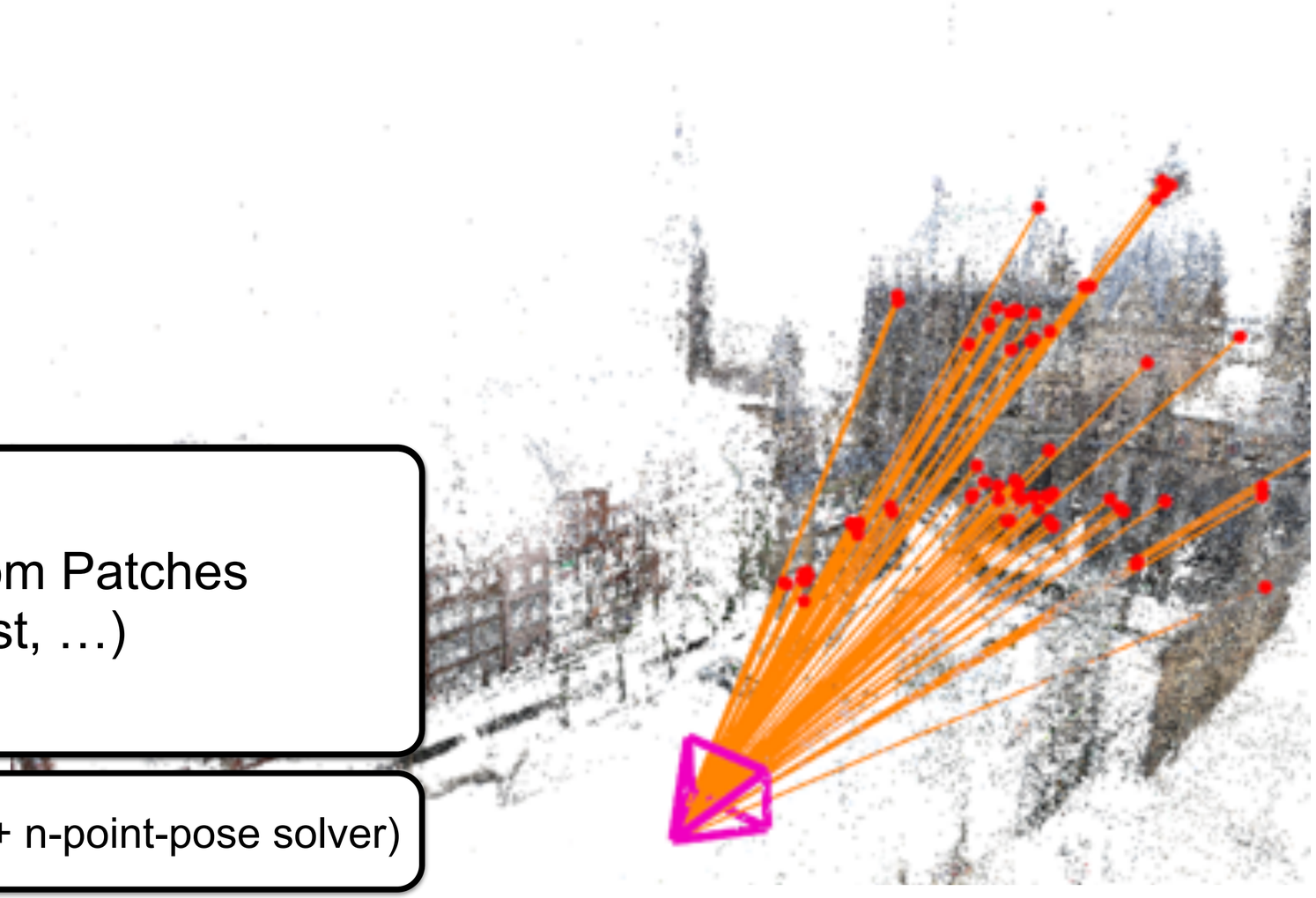
Feature Detection

Feature Description

Descriptor Matching for 2D-3D Matching

Estimate Camera Pose (RANSAC + n-point-pose solver)

Scene Coordinate Regression



Predict 2D-3D Matches from Patches
(CNN, Random Forest, …)

Estimate Camera Pose (RANSAC + n-point-pose solver)

Why Local Features?

- **Scalability**: Scene coordinate regression currently limited to smaller scenes [Brachmann & Rother, CVPR 2018]
- **Flexibility**: Easily add new data; compress scene by removing 3D points / compressing point descriptors [Li et al., ECCV 2010]
- **Generalization**: Local features can be trained to generalize to new scenes

Long-Term Visual Localization

Weakly Textured Scenes

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

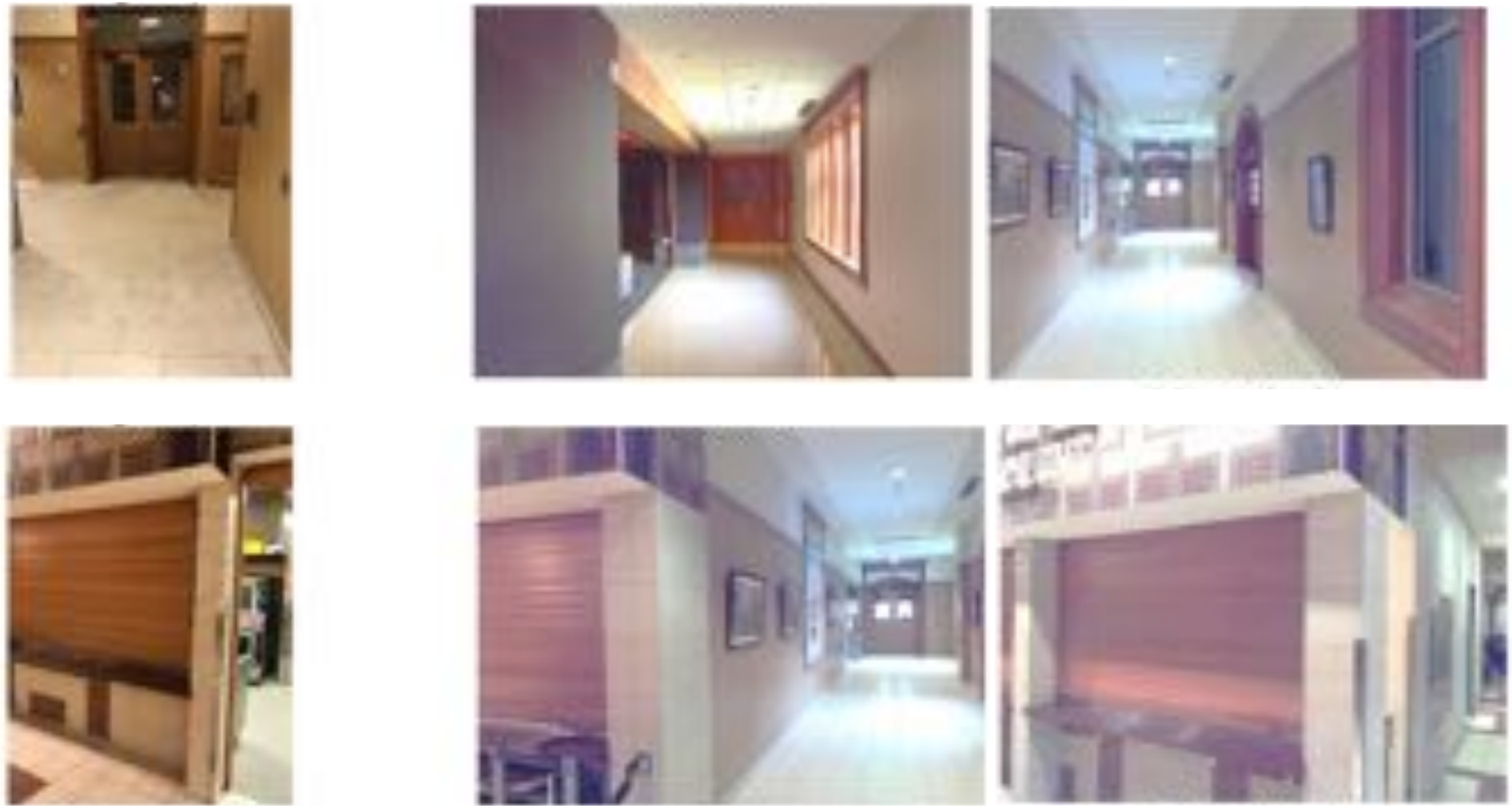
Complex Illumination

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

Changes in Viewpoint & Time

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

Repetitive Structures

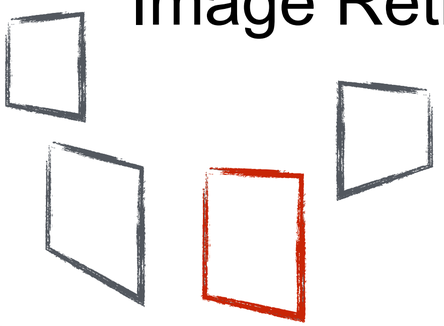


slide credit: Hajime Taira

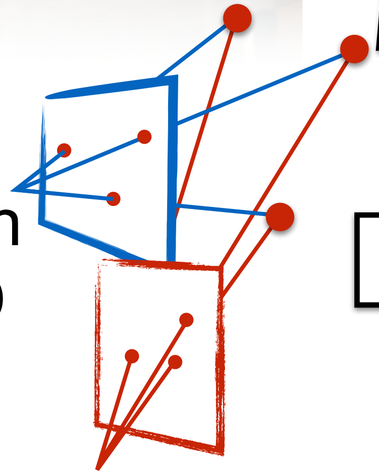
[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

Indoor Localization

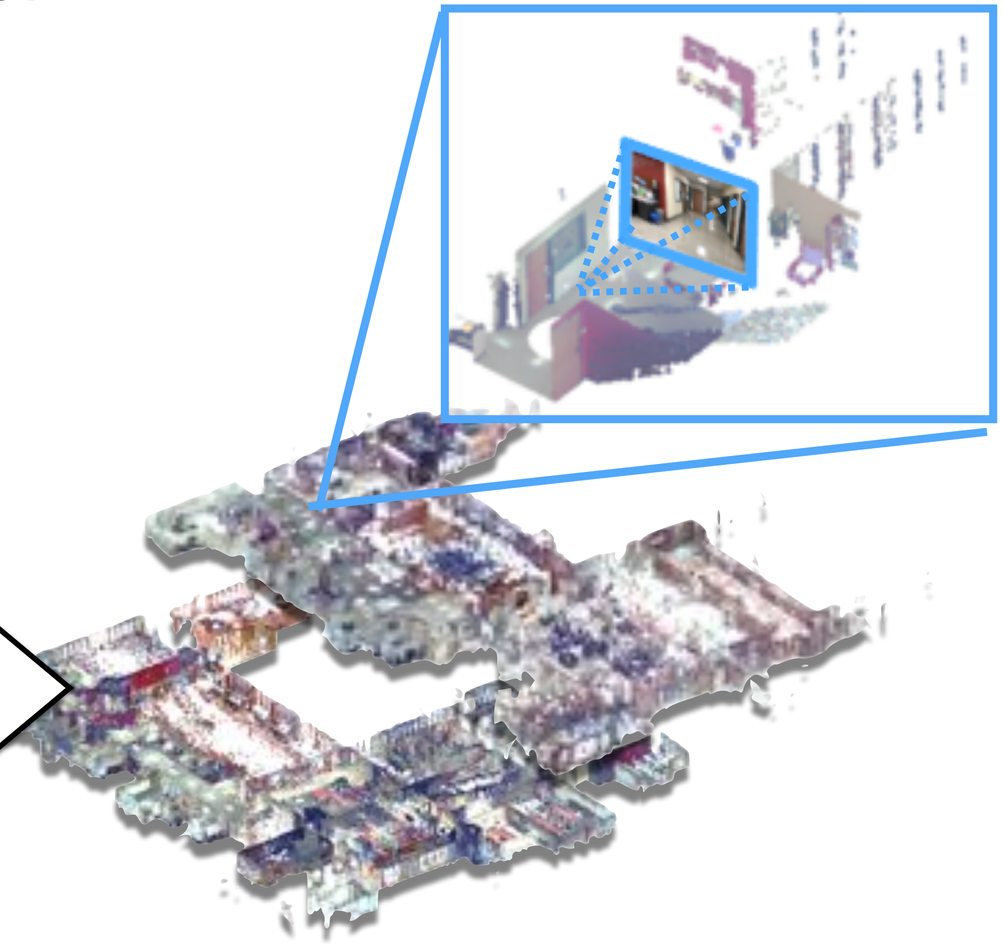
Image Retrieval



Matching &
Pose Estimation
against RGB-D
images



Pose
Verification



[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

Indoor Localization

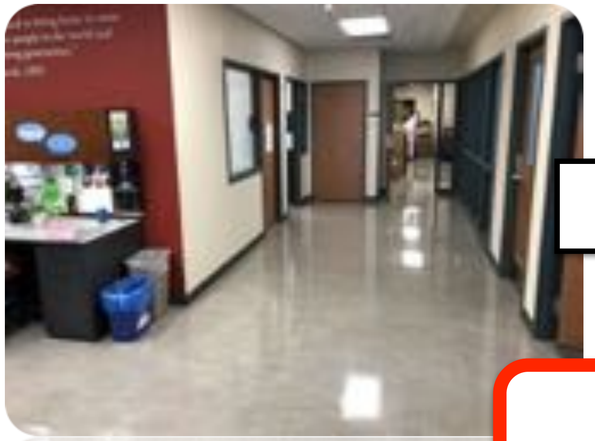
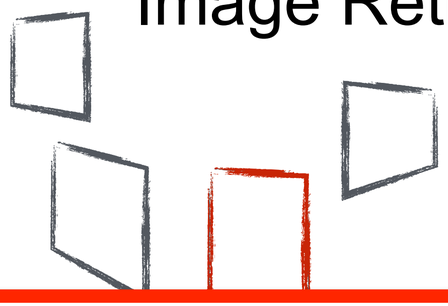
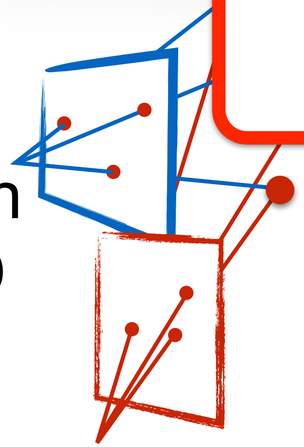


Image Retrieval



Twist: Perform each step using **dense information**

Matching & Pose Estimation against RGB-D images

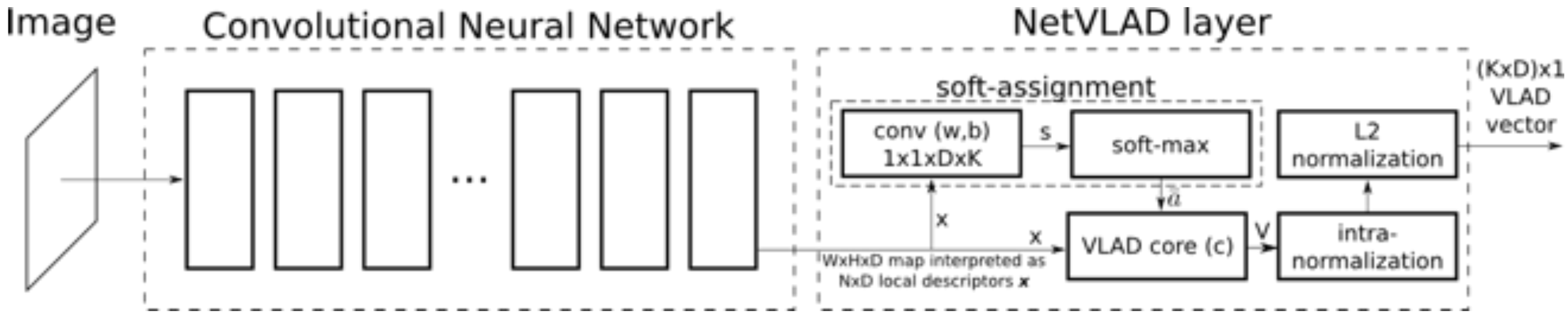


Pose Verification

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

“Dense” Indoor Localization

NetVLAD:



CNN, e.g., VGG

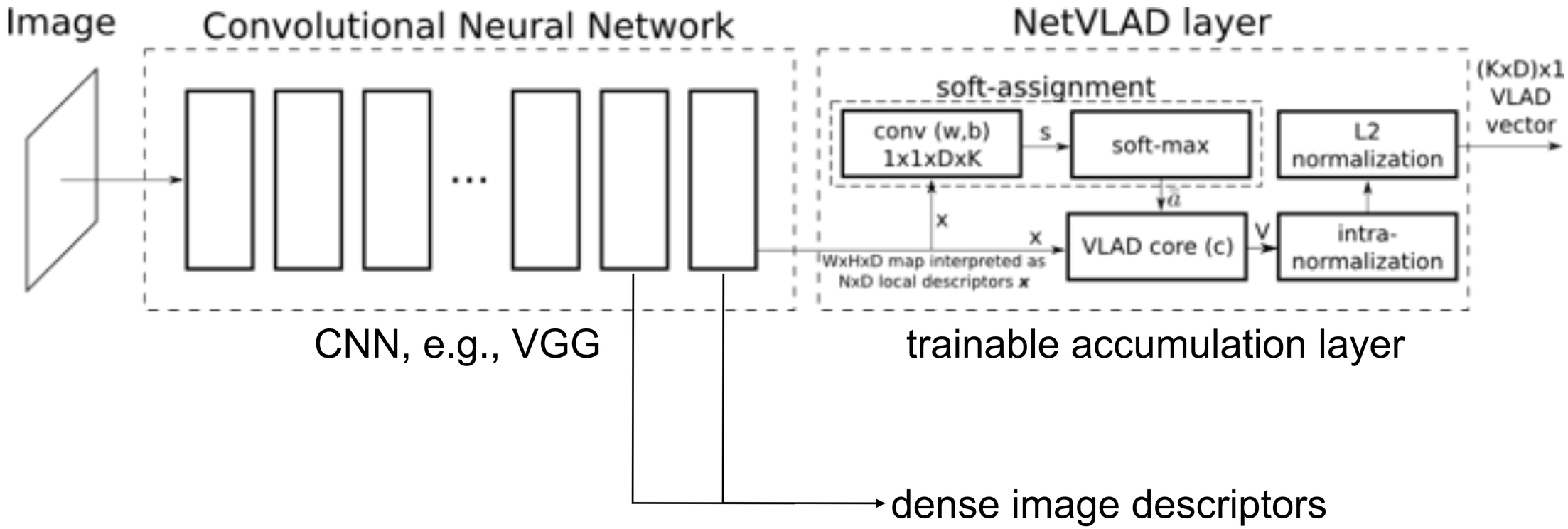
trainable accumulation layer

slide credit: Relja Arandjelović

[Arandjelović, Gronat, Torii, Pajdla, Sivic, NetVLAD: CNN architecture for weakly supervised place recognition. CVPR 2016]

“Dense” Indoor Localization

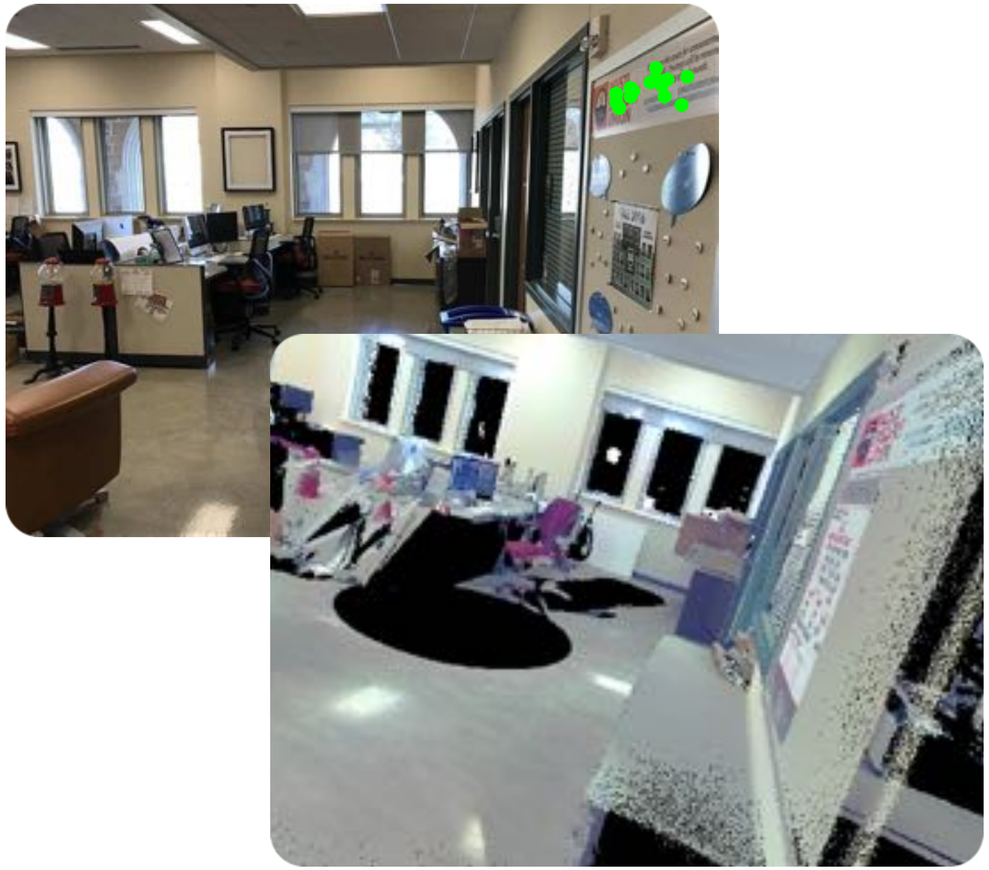
NetVLAD:



slide credit: Relja Arandjelović

[Arandjelović, Gronat, Torii, Pajdla, Sivic, NetVLAD: CNN architecture for weakly supervised place recognition. CVPR 2016]

Dense Feature Matching

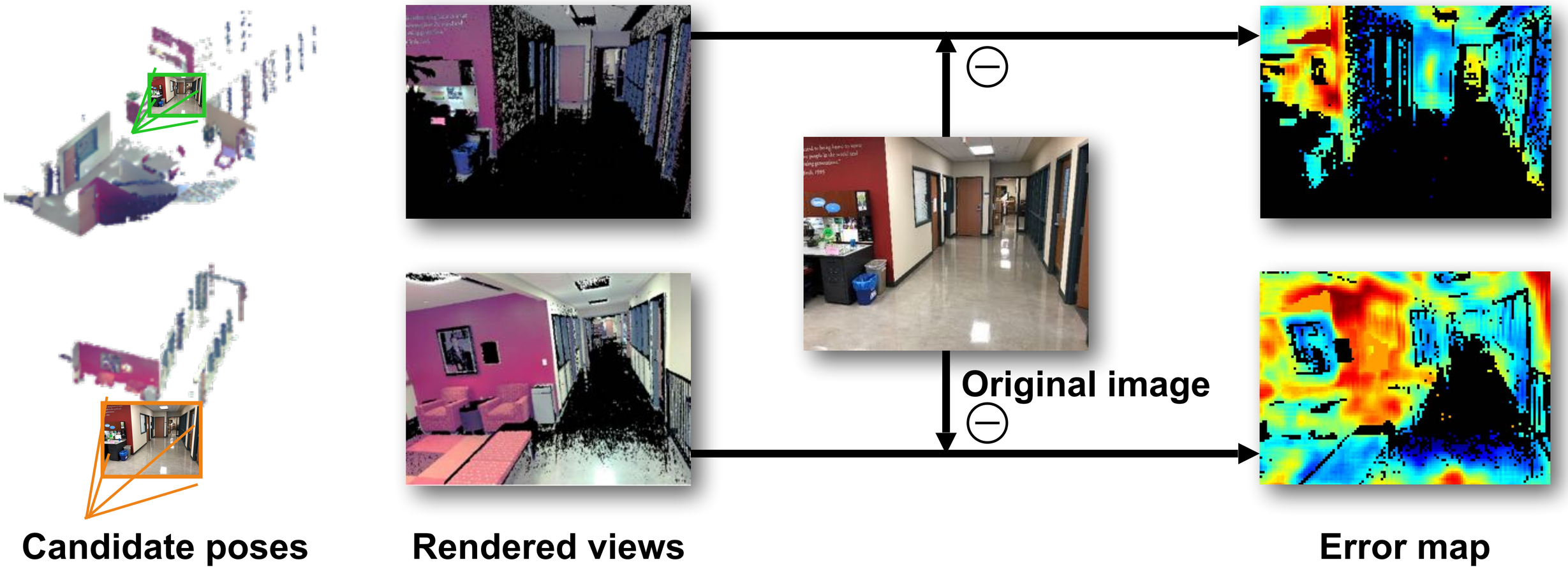


Sparse Matching

Dense Matching

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

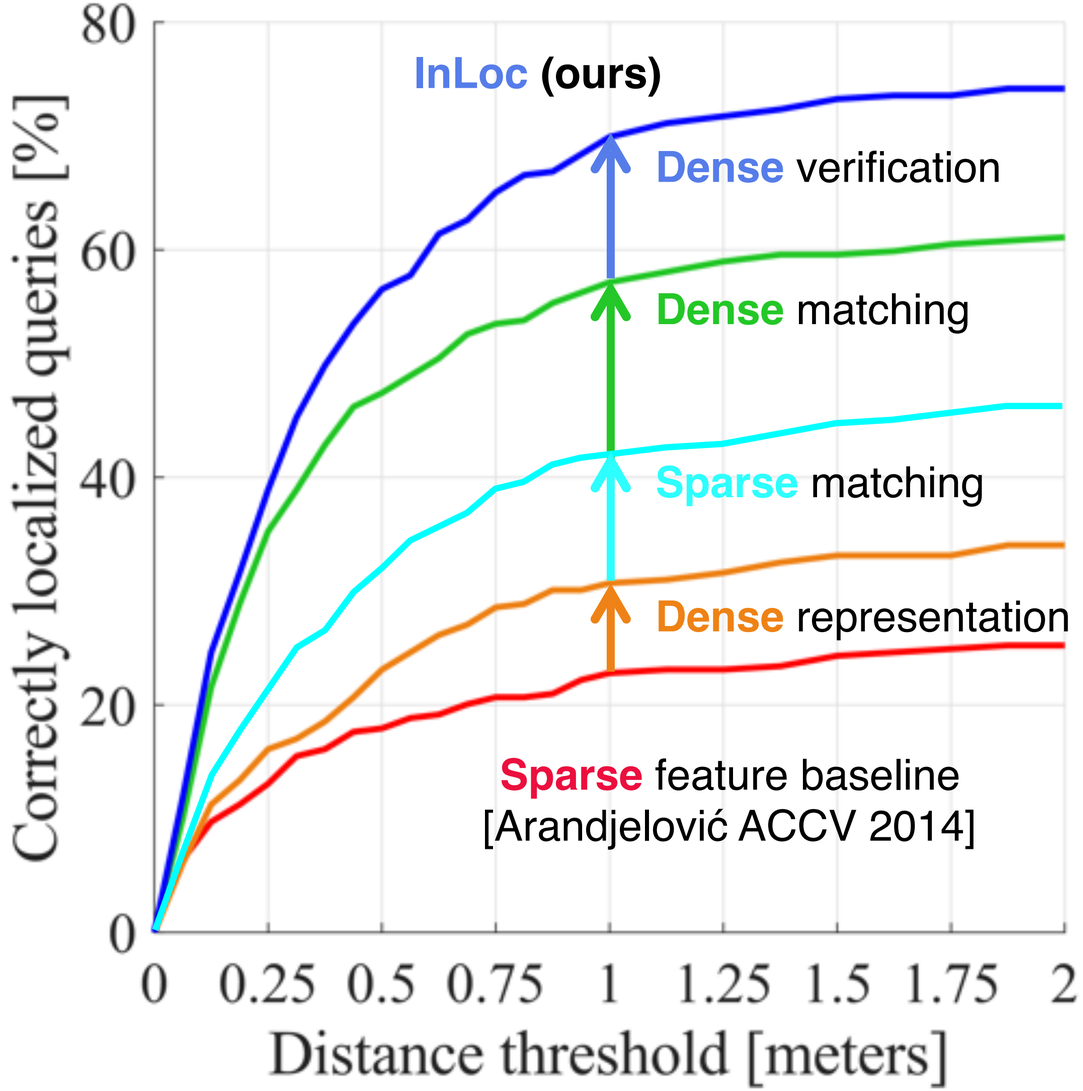
Dense Pose Verification



slide credit: Hajime Taira

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

“Dense” Indoor Localization



InLoc dataset

http://www.ok.sc.e.titech.ac.jp/INLOC/

slide credit: Hajime Taira

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

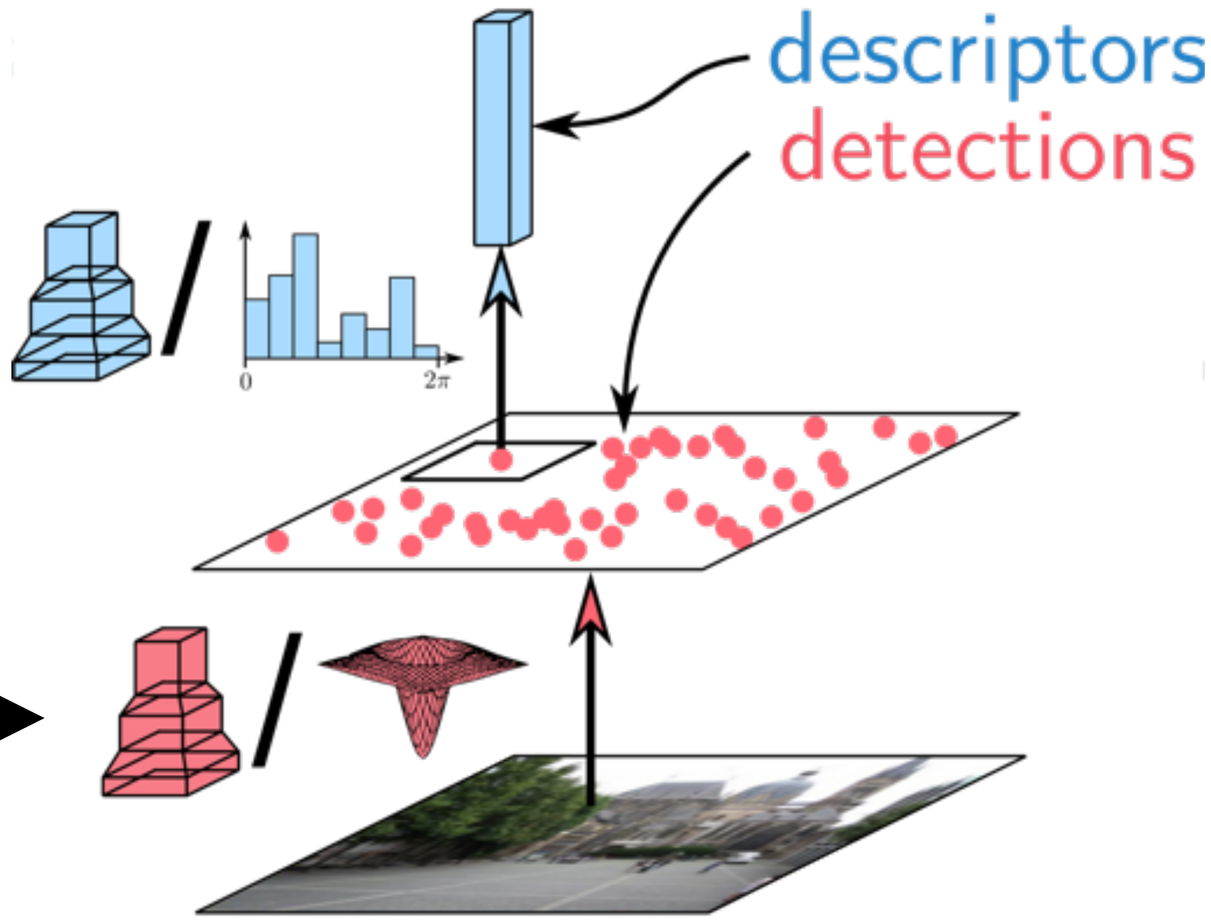
Main Lessons

- **Higher-Level Features**: Model higher-level structures rather than low-level information (gradients, …)
- **Skip Keypoint Detector**: Densely extract and match features
- **Limitations**: Run-time and memory
- **Do we really need dense features?**

[Taira, Okutomi, Sattler, Cimpoi, Pollefeys, Sivic, Pajdla, Torii, InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. CVPR 2018]

Classical Local Features

Detect-then-Describe:



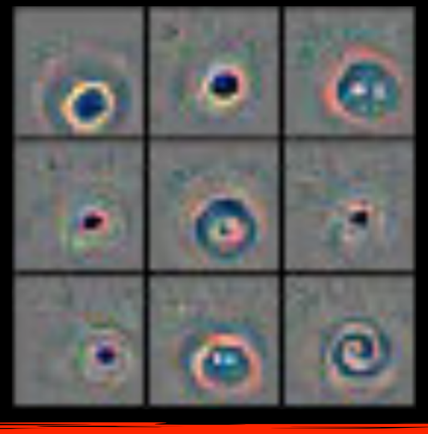
Efficient, looking at low-level structures / statistics

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

CNNs as Object Detectors

Low-Level Features

Mid-Level Features



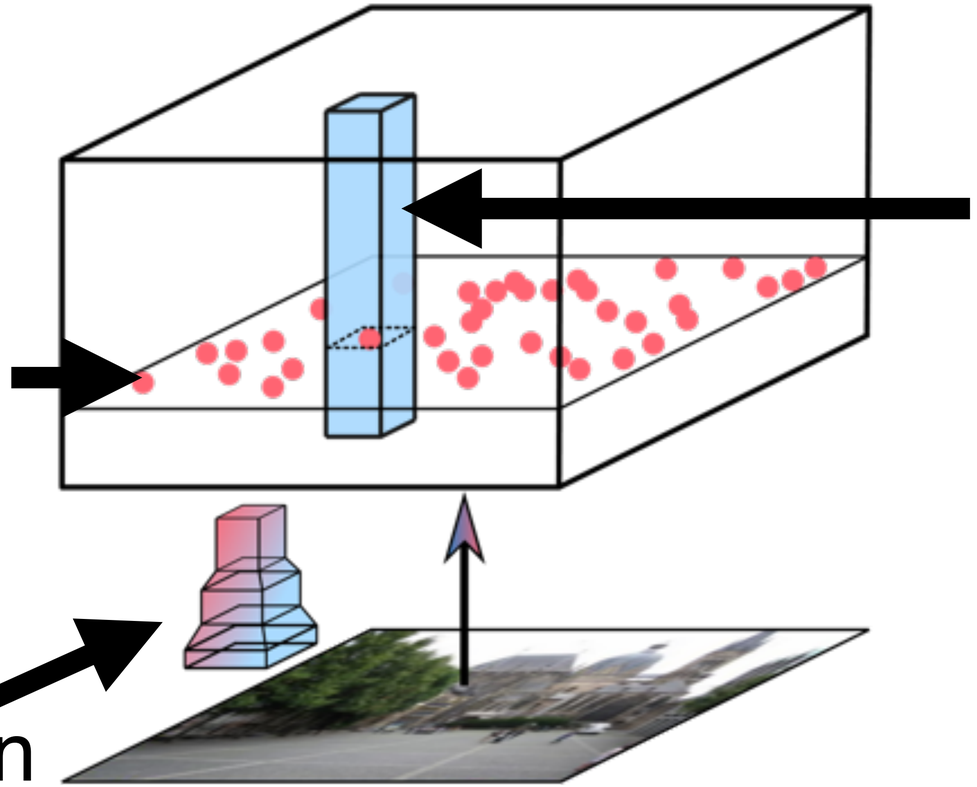
High-Level Features

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

Detect-And-Describe Approach

Descriptor ≈
“Objectness” scores

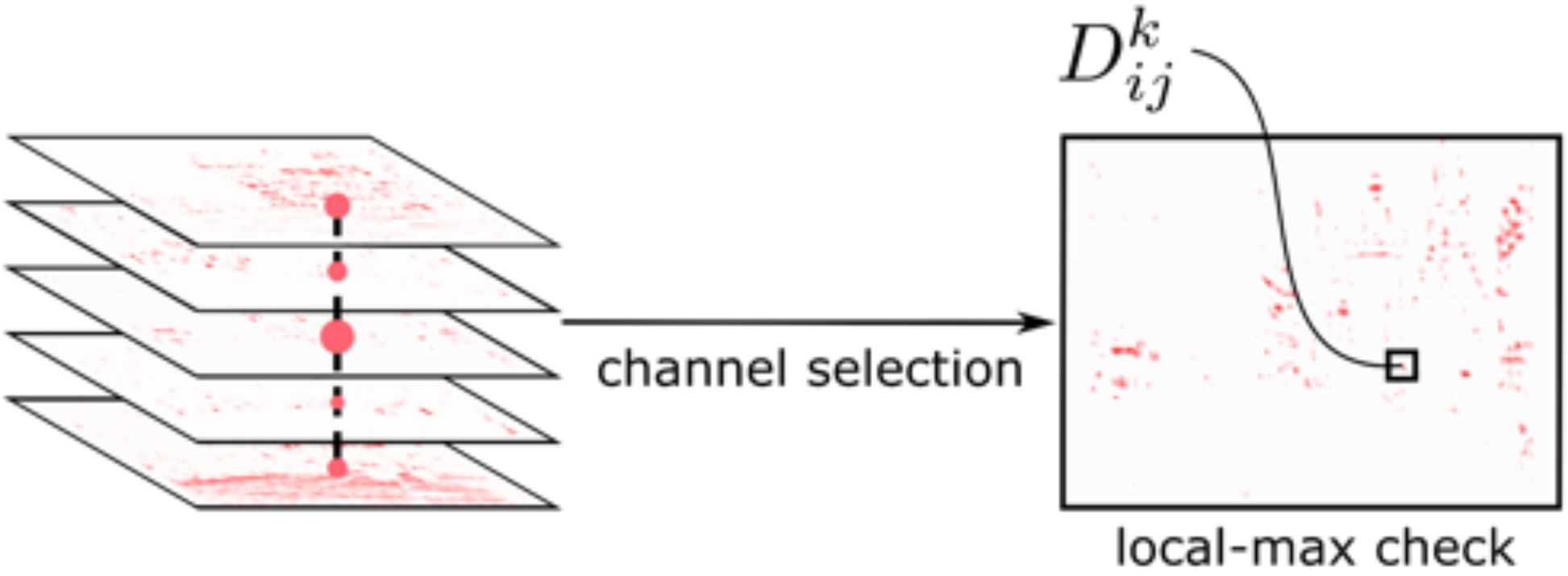
Local maxima ≈
Object detections



Same CNN for
detection & description

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

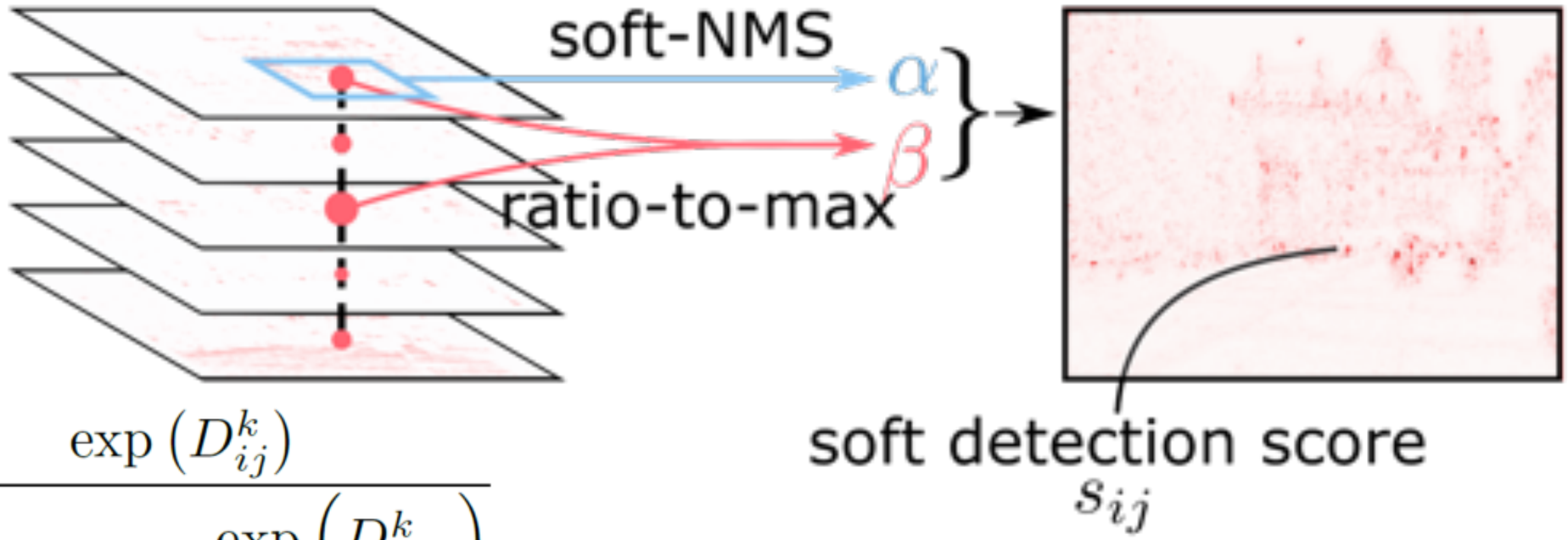
Hard Keypoint Detection



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

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

Soft Keypoint Detection



$$\alpha_{ij}^k = \frac{\exp\left(D_{ij}^k\right)}{\sum_{(i',j')\in\mathcal{N}(i,j)} \exp\left(D_{i'j'}^k\right)}$$

$$\beta_{ij}^k = D_{ij}^k / \max_t D_{ij}^t$$

$$\boxed{s_{ij} \propto \max_k (\alpha_{ij}^k \cdot \beta_{ij}^k)}$$ & triplet loss on descriptor

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

Aachen Day-Night Dataset

[Sattler, Maddern, Toft, Torii, Hammarstrand, Stenborg, Safari, Okutomi, Pollefeys, Sivic, Kahl, Pajdla, Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions, CVPR 2018]

The Good

Aachen Day-Night

Method	# Features	Correctly localized queries (%) 0.5m, 2°	1.0m, 5°	5.0m, 10°	10m, 25°
Upright RootSIFT	11.3K	36.7	54.1	72.5	81.6
DenseSfM	7.5K / 30K	39.8	60.2	84.7	99.0
HessAffNet + HardNet++	11.5K	39.8	61.2	77.6	88.8
SuperPoint	6.6K	42.8	57.1	75.5	86.7
DELF	11K	38.8	62.2	85.7	98.0
D2-Net SS (ours)	7K	41.8	66.3	85.7	98.0
D2-Net MS (ours)	11.4K	43.9	**67.3**	87.8	99.0
D2-Net SS Trained (ours)	14.5K	**44.9**	66.3	**88.8**	**100**
D2-Net MS Trained (ours)	19.3K	**44.9**	64.3	**88.8**	**100**

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

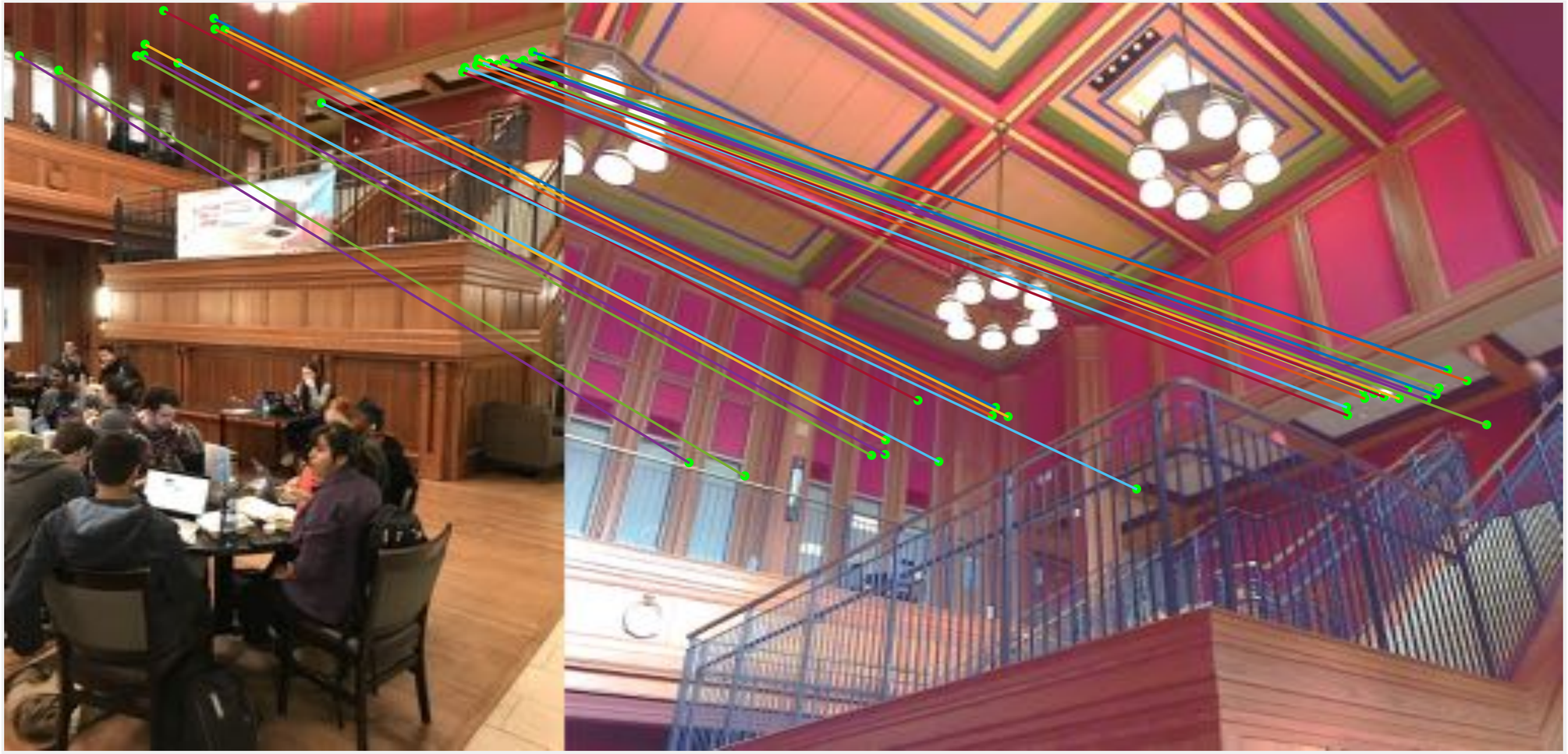
The Good

InLoc dataset

Method	Localized queries (%) 0.25m	0.5m	1.0m
Sparse PE - Aff. RootSIFT	21.3	32.2	44.1
Sparse PE - D2-Net MS (**ours**)	**35.0**	**48.6**	**62.6**
Dense PE	**35.0**	46.2	58.1
Sparse PE - Aff. RootSIFT + **Dense PV**	29.5	42.6	54.5
Sparse PE - D2-Net MS + **Dense PV** (**ours**)	38.0	54.1	65.4
Dense PE + **Dense PV** (= **InLoc**)	**41.0**	**56.5**	**69.9**
InLoc + D2-Net MS (**ours**)	**43.2**	**61.1**	**74.2**

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

The Good



[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

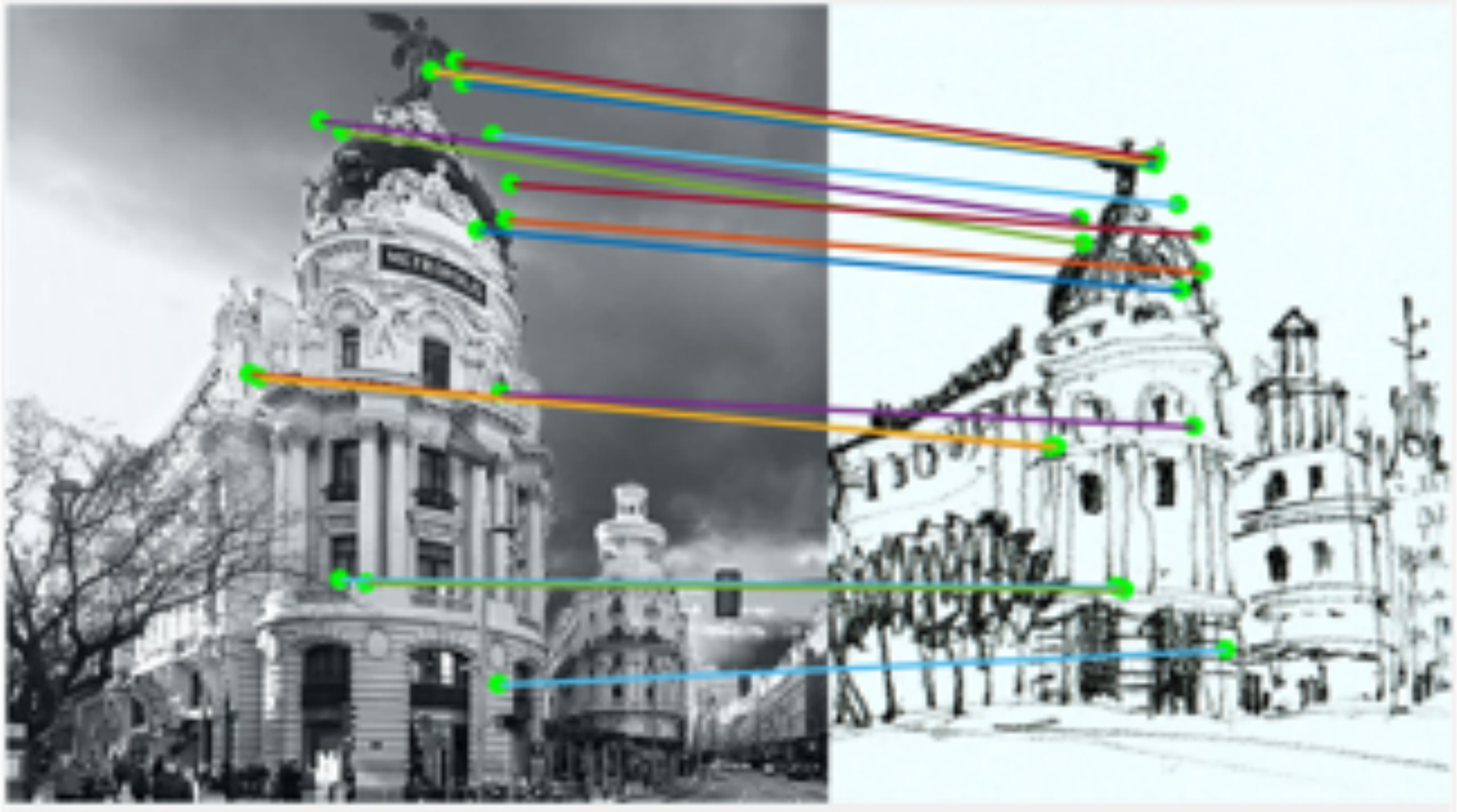
The Good

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

The Good

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

The Good

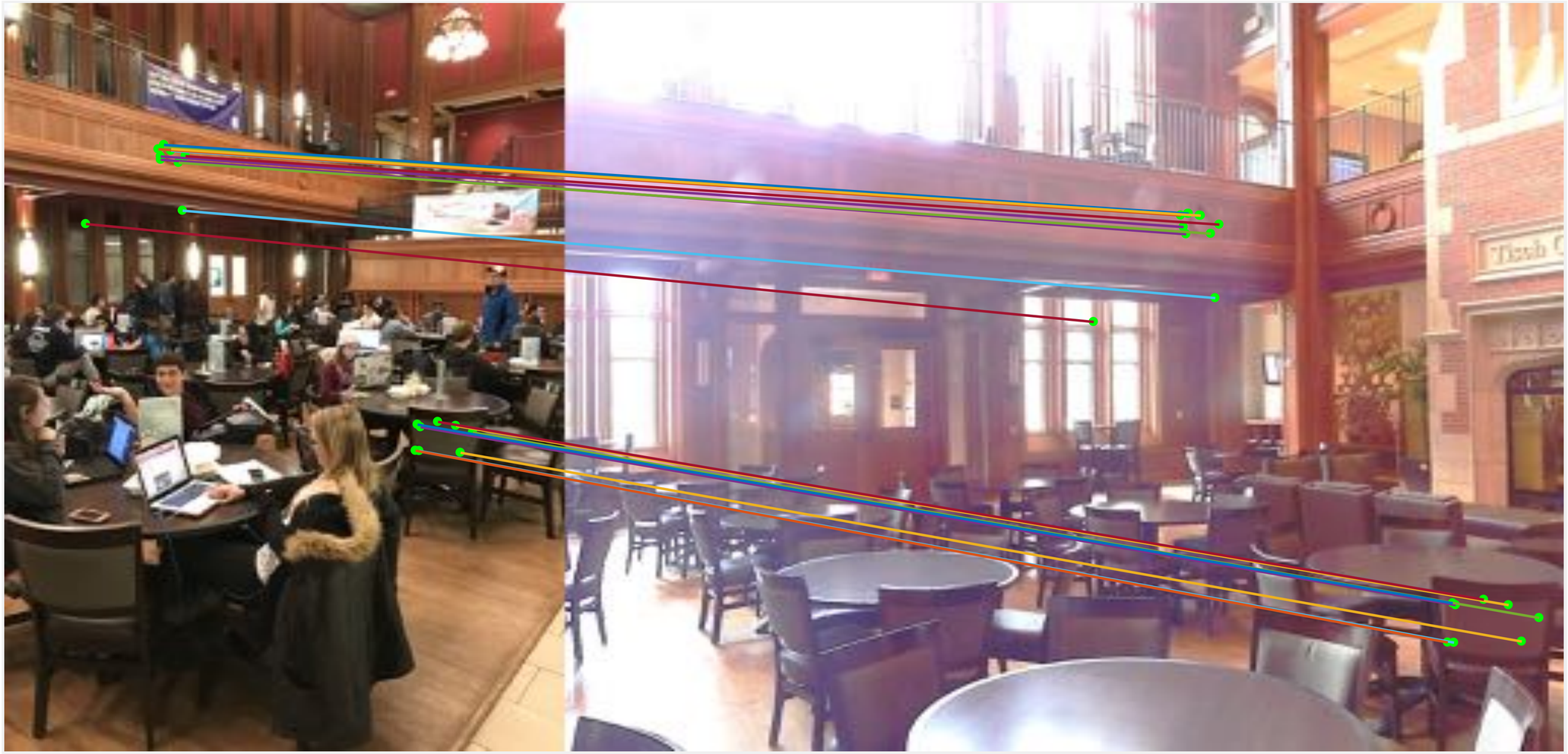


[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

The Ambiguous

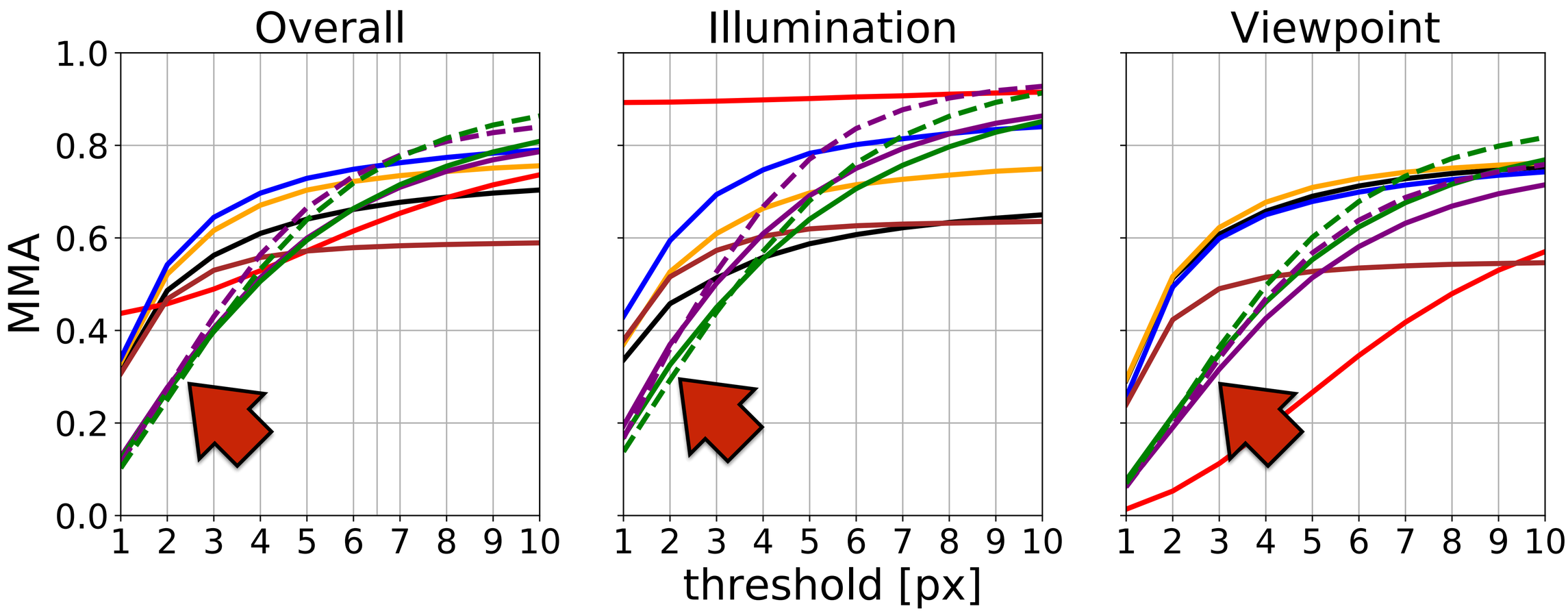
[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

The Ambiguous



[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

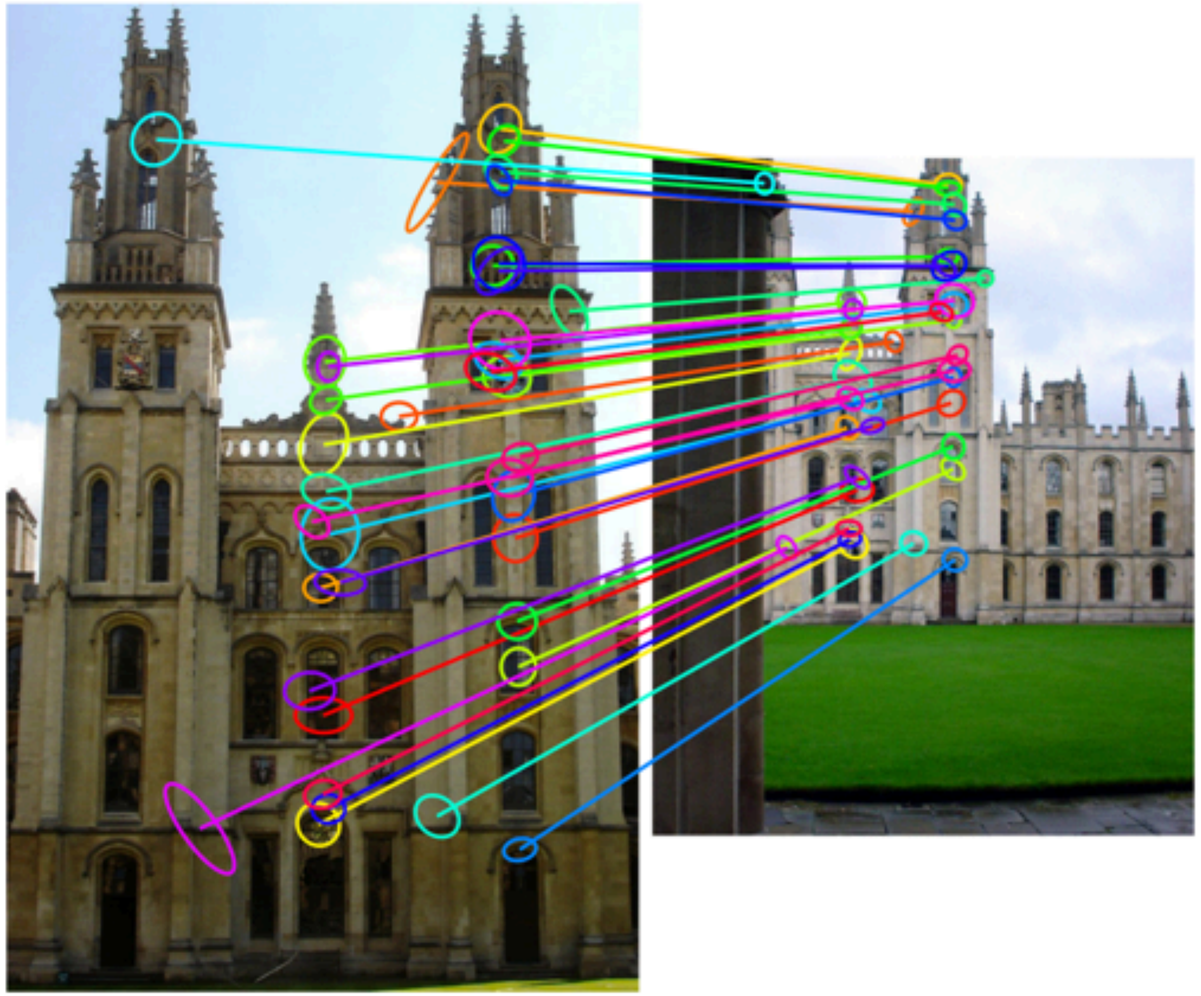
The Bad

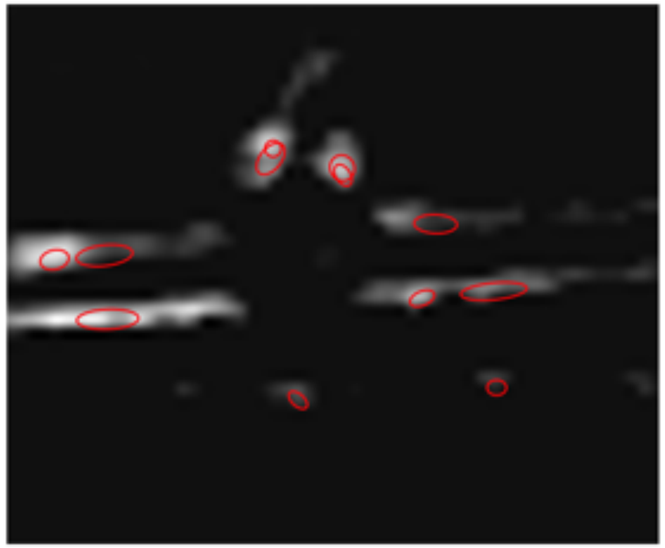


Our features are less well-localized (max-pooling hurts further)

[Dusmanu, Rocco, Pajdla, Pollefeys, Sivic, Torii, Sattler, D2-Net: A Trainable CNN for Joint Detection and Description of Local Features, CVPR 2019]

Similar Work on Visual Words / Image Retrieval





[Simeoni, Avrithis, Chum, Local Features and Visual Words Emerge in Activations, CVPR 2019]

Summary

- **Local features still relevant** for visual localization, CNNs for **pose regression do not work well**
- **Problems**:
  - Strong **changes** in **scene appearance & geometry**
  - **Low-level** vs. **higher-level structures**
  - Keypoint **detector** can be **limiting factor**
- **Joint detection and description** can help
  - … but detection accuracy is currently an issue
  - Code: **https://github.com/mihaidusmanu/d2-net**

visuallocalization.net

Monday afternoon, 1:30pm to 6pm, Hyatt Beacon B