Cross-weather-time, long term Visual Geo-Localization

CVPR 2021 tutorial on Cross-view and Cross-modal Visual Geo-Localization

Rakesh (Teddy) Kumar

Center for Vision Technologies
SRI International, Princeton, NJ, USA
Email: rakesh.kumar@sri.com

June 20th, 2021

Topics:

Visual Localization and Place Recognition

Cross weather/ time Visual Geo-localization

- Coarse Search
  - Feature Representations
    - Aggregated and Pooled representations
  - Dimensionality Reduction Techniques
  - Learning to Retrieve
    - Siamese Networks, Triplet loss, Ranked List Loss
  - Semantic Networks and Attention
- Fine Geo-localization
  - Multi-headed networks for learning local and global features simultaneously
  - SuperGlue, Graph based multi-attention matching using context
- Concluding remarks

Papers covered

Coarse Search

1. NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016, https://arxiv.org/abs/1511.07247

2. Fine-tuning CNN Image Retrieval with No Human Annotation, Radenović F., Tolias G., Chum O., TPAMI 2018, https://arxiv.org/abs/1711.02512

3. Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020, https://arxiv.org/abs/2007.12163

4. Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization, Zachary Seymour, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, Rakesh Kumar, BMVC 2019, https://arxiv.org/abs/1812.03402

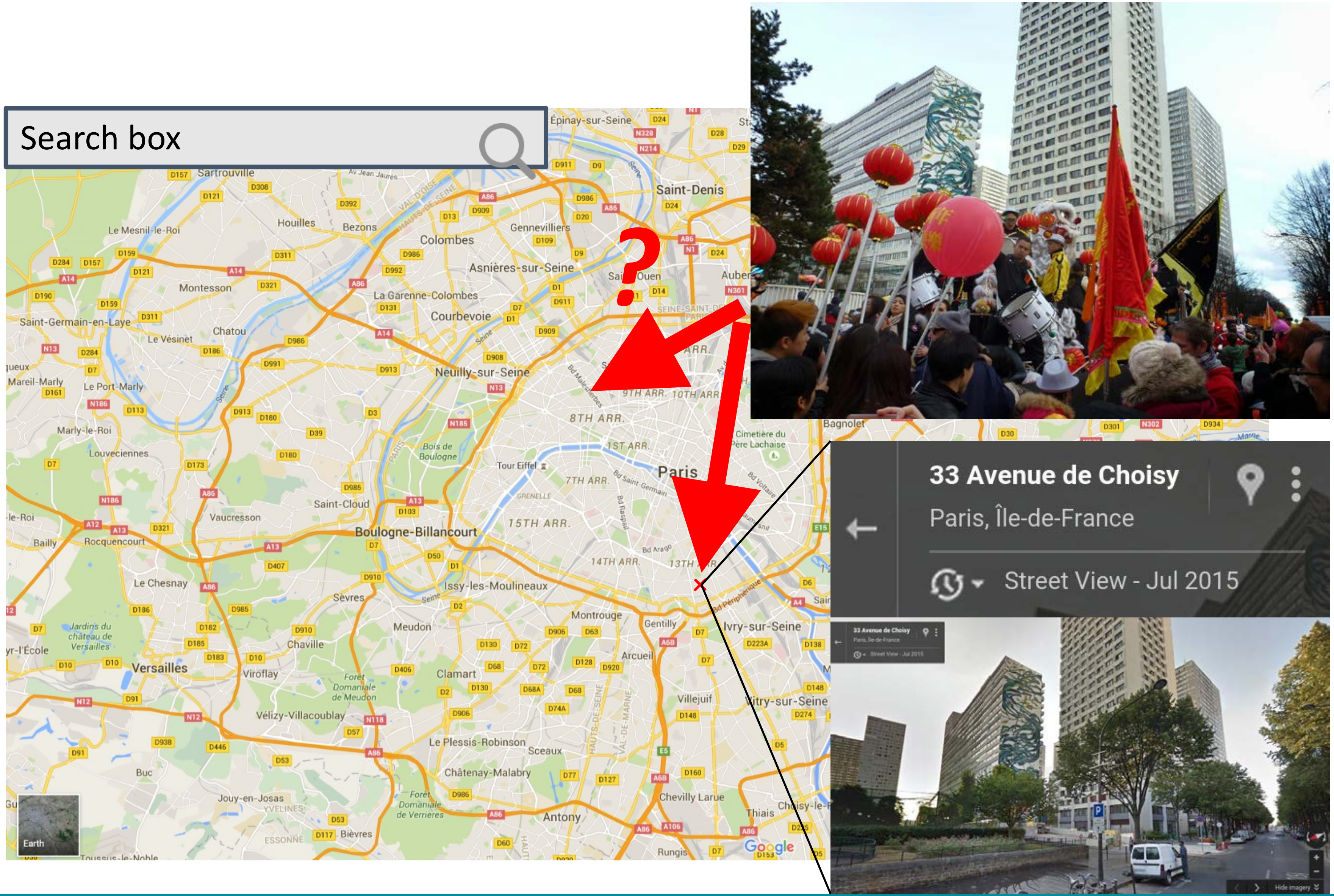
Fine Geo-localization and end-to-end solutions

5. From Coarse to Fine: Robust Hierarchical Localization at Large Scale, Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, Marcin Dymczyk, CVPR 2019, https://arxiv.org/abs/1812.03506

6. SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020. https://arxiv.org/abs/1911.11763, https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

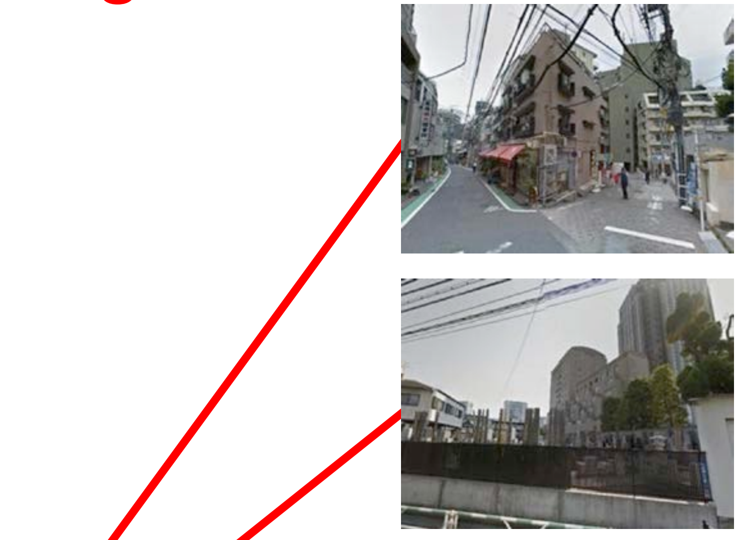
7. Robust Image Retrieval-based Visual Localization using Kapture https://arxiv.org/abs/2007.13867

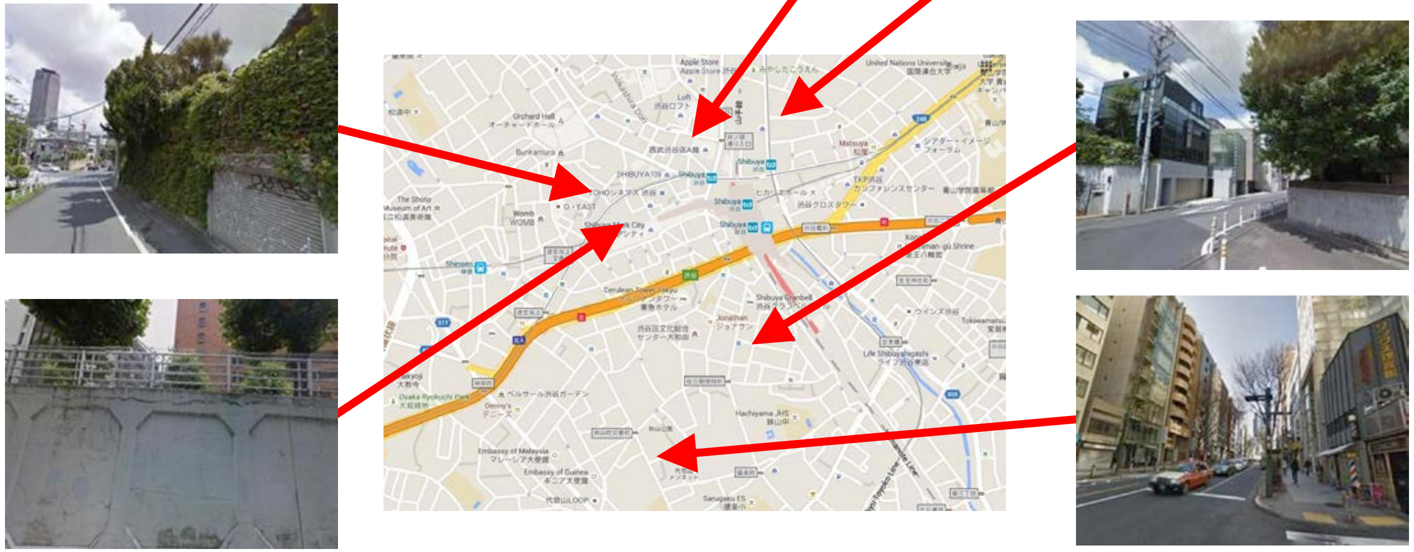
Image based Geo-Localization



Visual instance recognition

Represent the world by a set of geotagged images

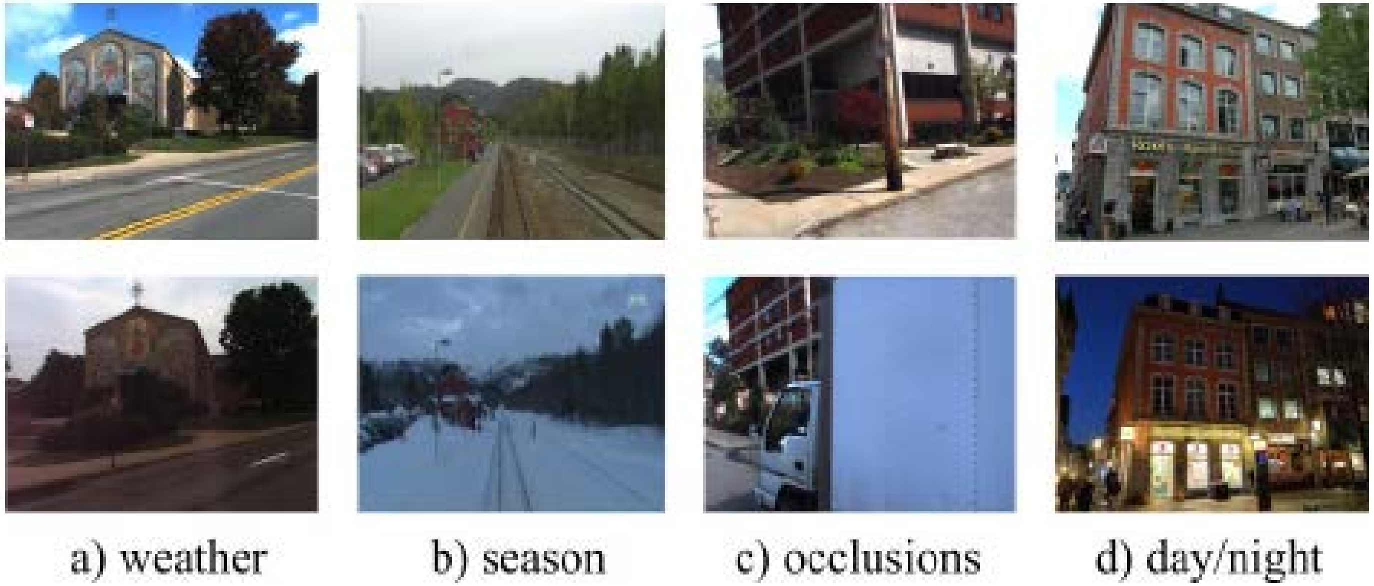




NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

Why is it a difficult problem

- Lighting changes: Different time of day / year
- Changes in camera viewpoint
- Occluders and ambiguous objects: Trees, cars, pavement…
- Big data: World-scale localization



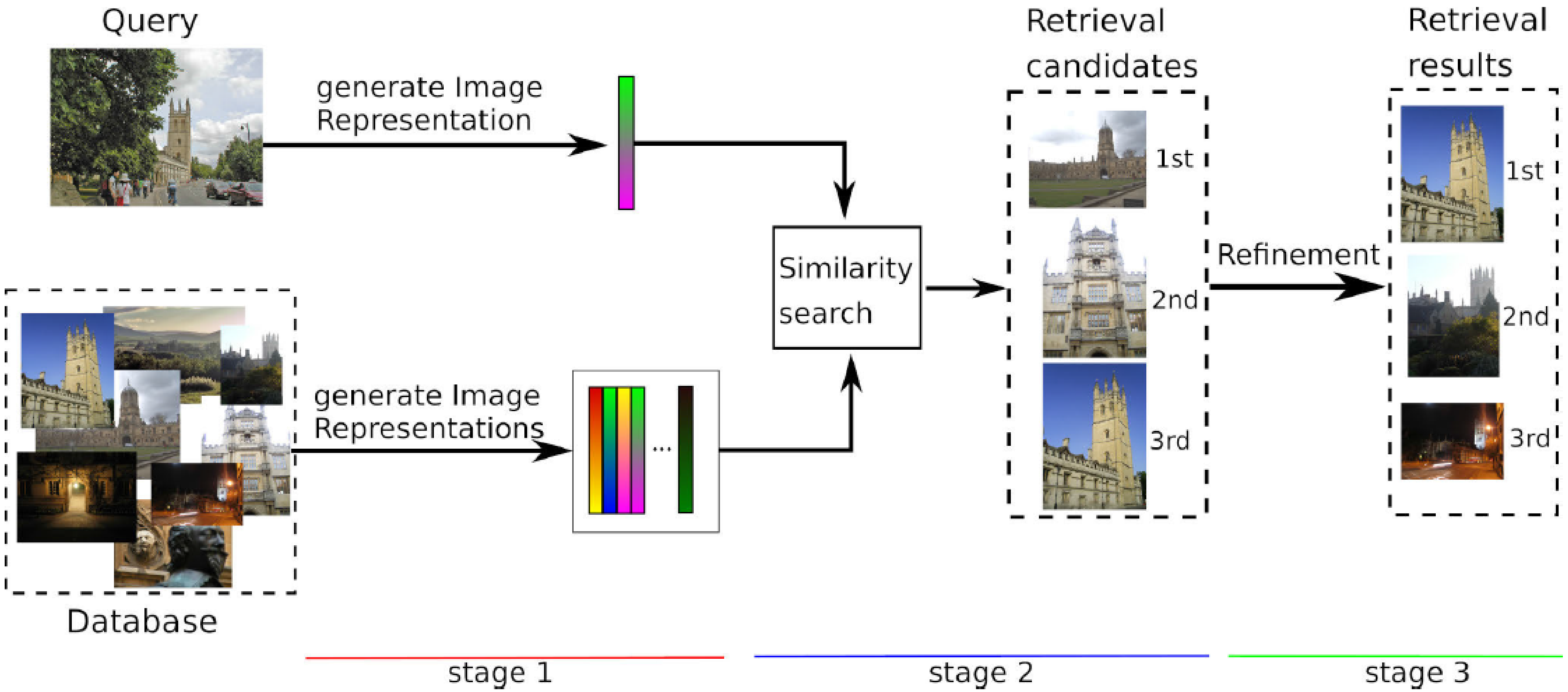
Examples of challenging conditions

Datasets for Visual Place Recognition

TABLE 1. Summary of commonly used datasets in VPR. Among the changing conditions, D/N stands for Day/Night, W stands for Weather, and S stands for Season. The column denoted as 3D indicates if the dataset includes 3D models.

Dataset	Date	Scene	Scale	# Images	Changing Conditions			3D	Place
					D/N	W	S		
Oxford [116]	2007	Urban	City	~5k					Label
Paris [122]	2008	Urban	City	~6k					Label
Holidays [118]	2008	Outdoor	World	~2k					Label
Eynsham [21]	2009	Urban	City	~70k					GPS
St. Lucia [240], [241]	2010	Urban	City	~66k					GPS
European Cities 50k [22]	2010	Urban	Continent	~50k					Label
Geotagged StreetView [23]	2010	Urban	City	~17k					GPS
Rome 16k [242]	2010	Urban	City	~16k				✓	Pose
Dubrovnik 6k [242]	2010	Urban	City	~6.8k				✓	Pose
San Francisco [243]	2011	Urban	City	~1.06M					GPS
Alderley [45]	2012	Urban	City	~31k	✓	✓			GPS
7 Scenes [244]	2013	Indoor	Building	~43k				✓	Pose
Nordland [155]	2013	Outdoor	Region	~143k			✓		GPS
Google StreetView 62k [114]	2014	Urban	City	~62k					GPS
Freiburg Across Seasons [192], [245]	2014	Urban	City	~43k			✓		GPS
Cambridge Landmarks [215]	2015	Urban	City	~10.8k				✓	Pose
Paris500k [246]	2015	Urban	City	~504k					Label
Pittsburgh [117]	2015	Urban	City	~278k					GPS
Landmarks-full [80], [125]	2016	Urban	World	~192k					Label
NCLT [247]	2016	Outdoor + Indoor	Campus	~3.8M		✓	✓	✓	Pose
Oxford Robotcar [248]	2017	Urban	City	~20M	✓	✓	✓		GPS
SPED [190]	2017	Outdoor	World	~1.3M	✓	✓	✓		Label
Google-Landmarks [38], [81]	2017	Outdoor	World	~1.2M					GPS
$\mathcal{R}$Oxford [129]	2018	Urban	City	~5k					Label
$\mathcal{R}$Paris [129]	2018	Urban	City	~6k					Label
Tokyo 24/7 [121]	2018	Urban	City	~ 2.8M	✓				GPS
Aachen Day/Night [151], [153], [154]	2018	Urban	City	~7.6k	✓			✓	Pose
RobotCar Seasons [151]	2018	Urban	City	~31k	✓	✓	✓	✓	Pose
CMU Seasons [151], [152]	2018	Urban	City	~116k	✓	✓	✓	✓	Pose
TokyoTM [69]	2018	Urban	City	~190k	✓				GPS
InLoc Dataset [119], [209]	2018	Indoor	Building	~10k				✓	Pose
TB Places v2 [249], [250]	2019	Garden	City	~59k					Label
San Francisco Revisited [214]	2019	Urban	City	~790k				✓	Pose
WorldCities [30]	2019	Urban	City	~300k					GPS
Google-Landmarks v2 [251]	2020	Outdoor + Indoor	World	~4.2M					GPS
Mapillary SLS [252]	2020	Urban	World	~1.68M	✓	✓	✓		GPS

Feature Representation for Image Retrieval

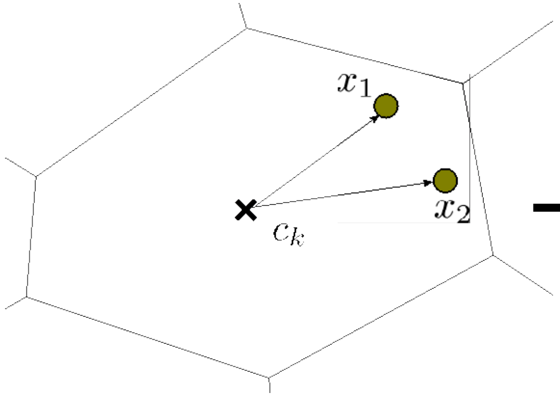


Visual place recognition is commonly formulated as an image retrieval problem. The known places are collected in a database and a new image to be localized is called query. The place retrieval is performed in three logical stages.

1) In the first stage, vector representations are generated for the query and the database images. From a practical perspective, the representation of the query is computed online, whereas the representations of the database images are computed offline.
2) The representation of the query is compared to those of the database images, to find the most similar ones (here only the top 3 are shown).
3) The best results of the comparison are further refined with post-processing techniques (here only the top3 are shown).

From: C. Masone and B. Caputo, "A Survey on Deep Visual Place Recognition," in IEEE Access, vol. 9, pp. 19516-19547, 2021, doi: 10.1109/ACCESS.2021.3054937.

Classical architecture versus deep learning architecture for place recognition



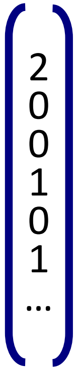


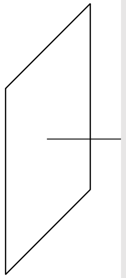
Image I

Extract local features (SIFT)

Aggregate (BoW, **VLAD**, FV)

f(I)

Image

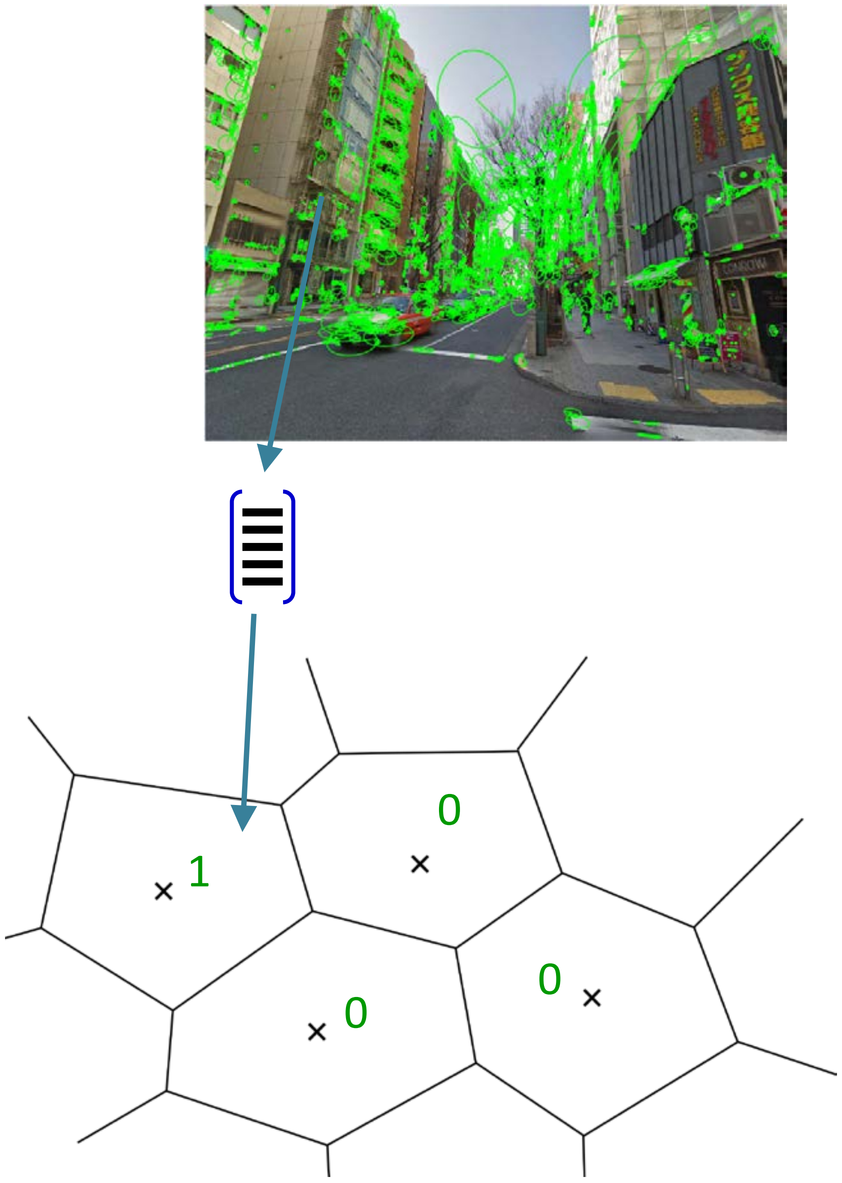


Make it trainable end-to-end

NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

Review: Pooling local descriptors - Bag-of-Words (BoW)

$$\overbrace{a_k(x_i)}^{\text{0/1 assignment of desc. } i \text{ to cluster } k}$$

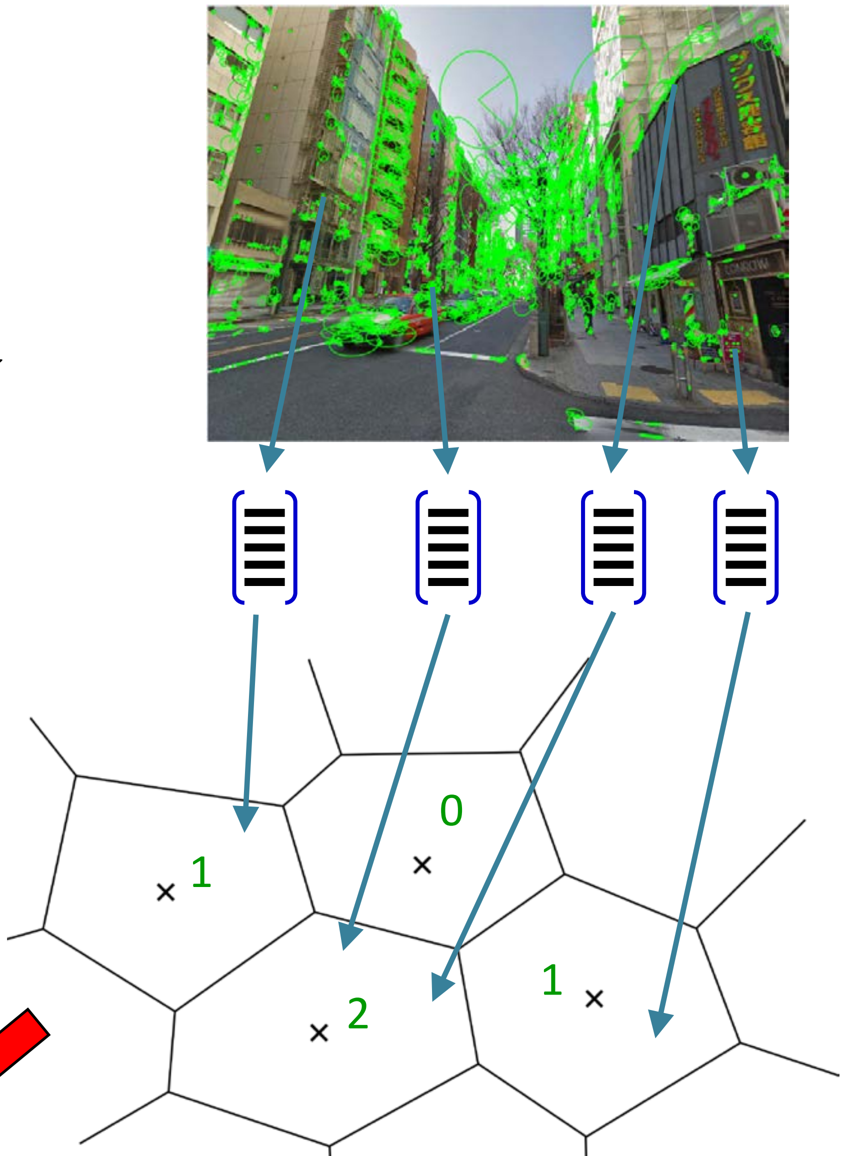


[Sivic and Zisserman 03]

Review: Pooling local descriptors - Bag-of-Words (BoW)

$$B(k) = \underbrace{\sum_{i=1}^{N}}_{\text{Sum over all N descriptors in the image}} \overbrace{a_k(x_i)}^{\text{0/1 assignment of desc. } i \text{ to cluster } k}$$

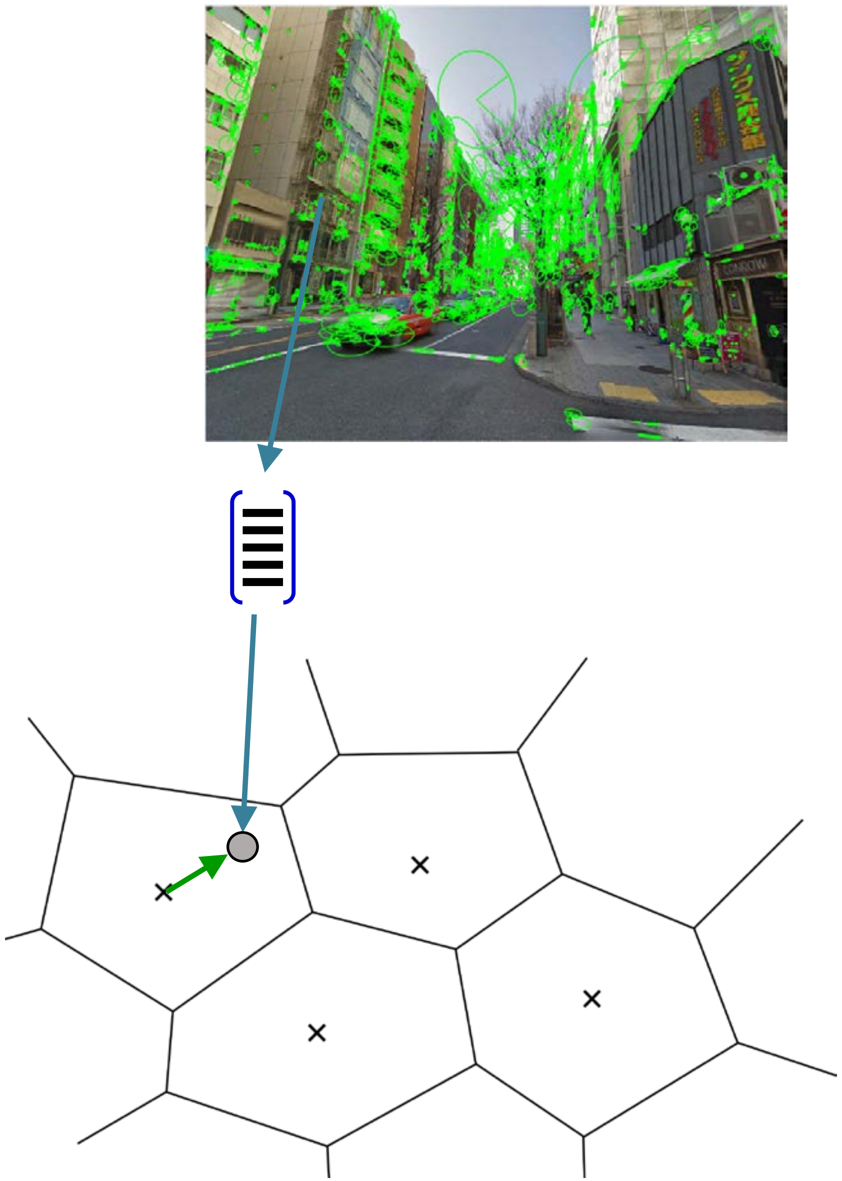
B = [1, 0, 2, 1, …]



[Sivic and Zisserman 03]

Review: Vector of Locally Aggregated Descriptors (VLAD)

$$\overbrace{a_k(x_i)}^{\text{0/1 assignment of desc. } i \text{ to cluster } k}\underbrace{(x_i - c_k)}_{\text{Residual vector}}$$

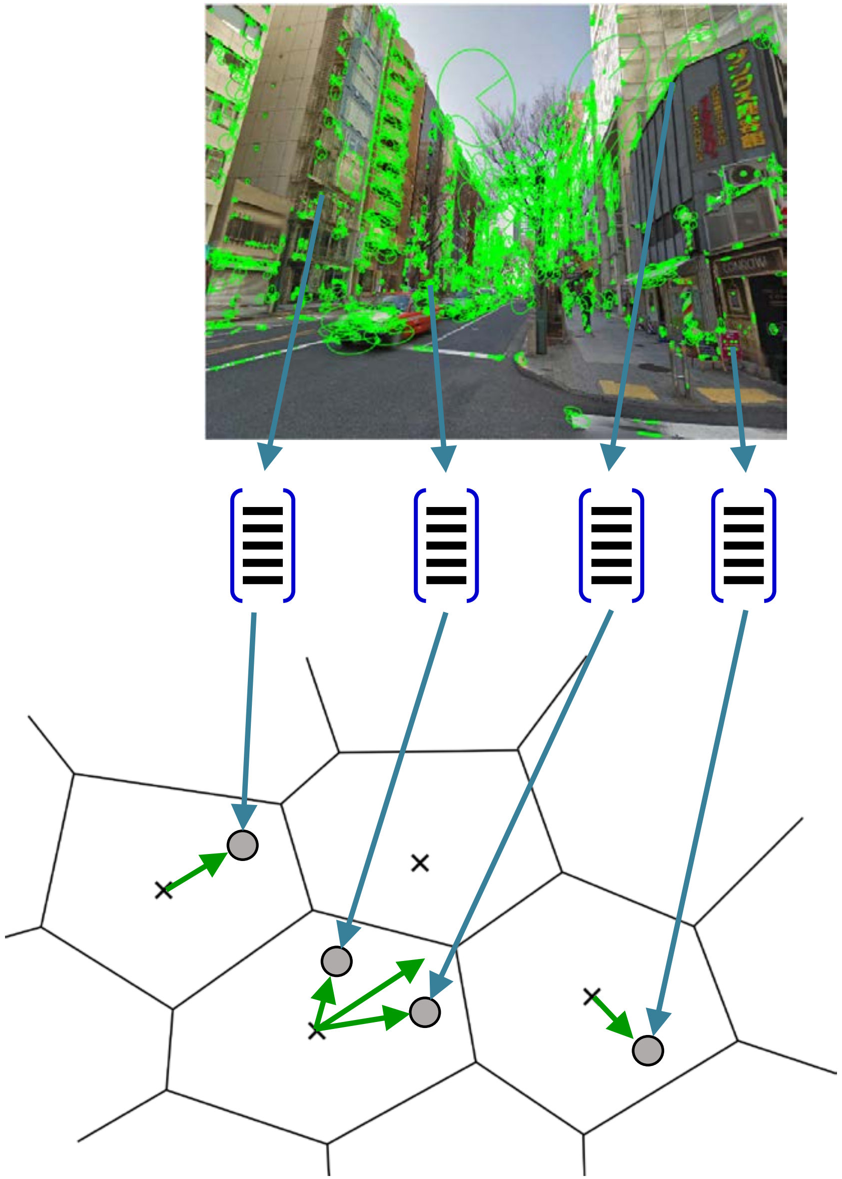


[Jégou et al. 10]

Review: Vector of Locally Aggregated Descriptors (VLAD)

$$V(:,k) = \underbrace{\sum_{i=1}^{N}}_{\text{Sum over all N descriptors in the image}} \overbrace{a_k(x_i)}^{\text{0/1 assignment of desc. } i \text{ to cluster } k} \underbrace{(x_i - c_k)}_{\text{Residual vector}}$$

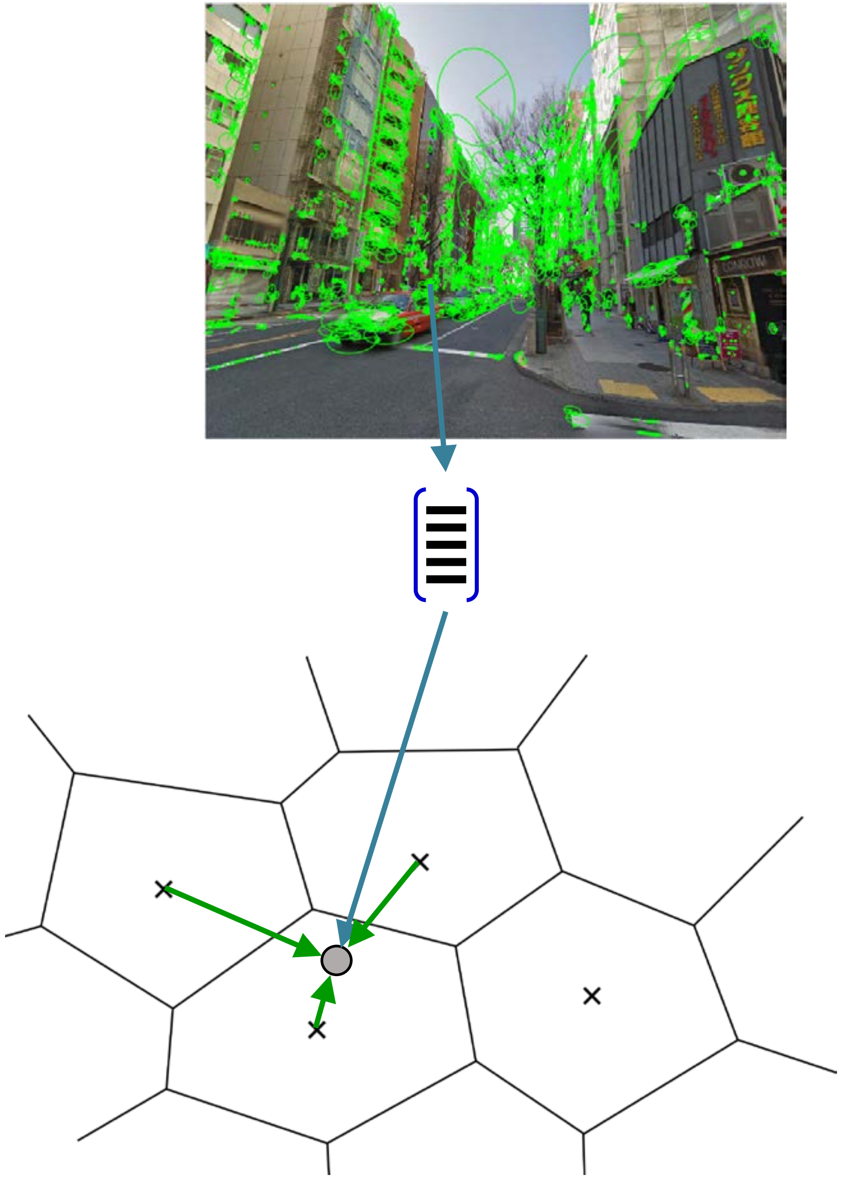
V = [↗, . . , ↗, ↘, …]



[Jégou et al. 10]

NetVLAD: Trainable pooling layer

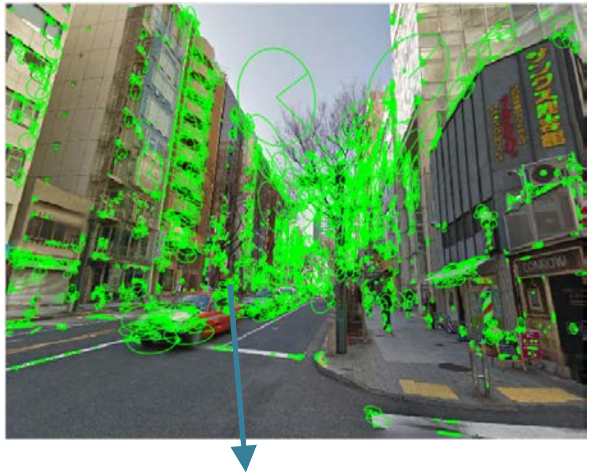
$$V(:,k) = \sum_{i=1}^{N} \overbrace{a_k(x_i)}^{\text{0/1 assignment of desc. } i \text{ to cluster } k}(x_i - c_k)$$



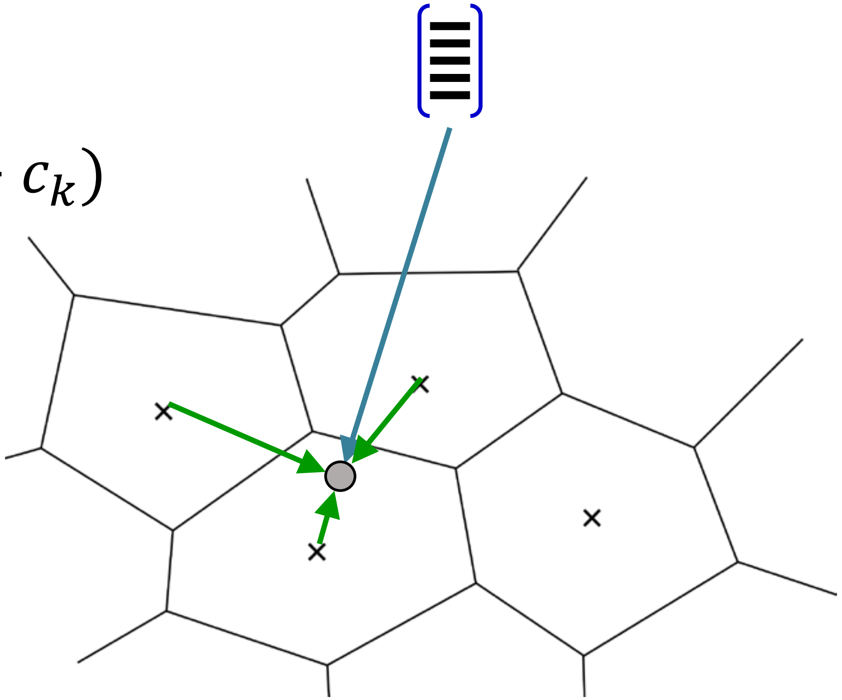
Replace hard-assignment of descriptors to clusters with soft-assignment

NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

NetVLAD: Trainable pooling layer



$$V(:,k) = \sum_{i=1}^{N} \overbrace{\boxed{\frac{e^{w_k^T x_i + b_k}}{\sum_{k'} e^{w_{k'}^T x_i + b_{k'}}}}}^{\text{soft assignment of desc. } i \text{ to cluster } k} (x_i - c_k)$$



NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

NetVlad: Mimic the state-of-the-art architecture

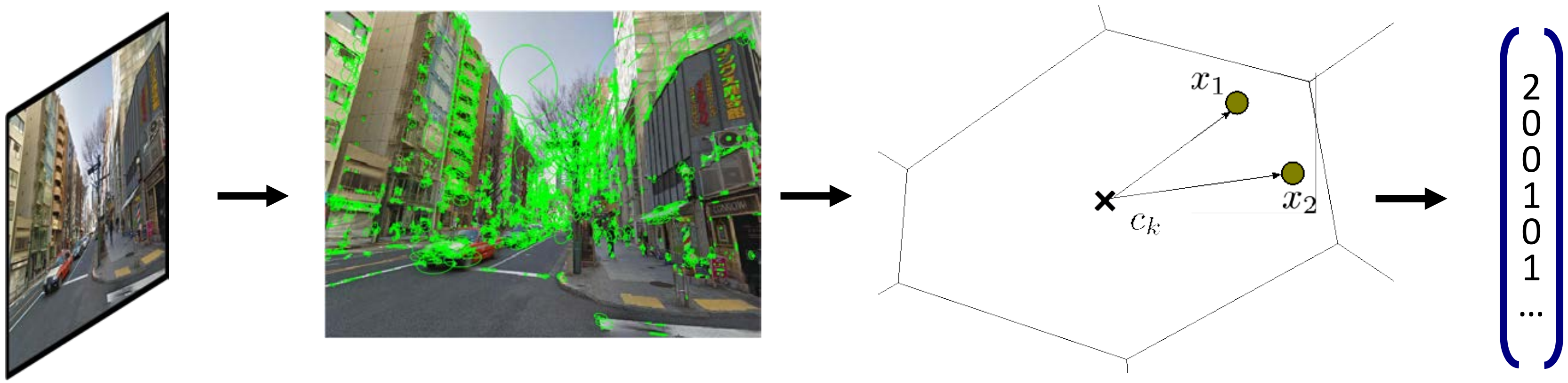


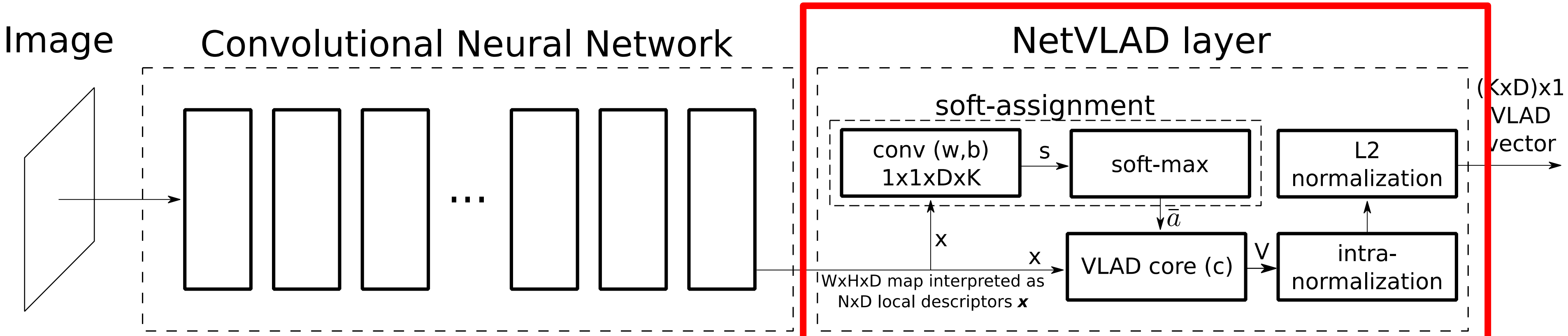
Image I

Extract local features (SIFT)

Aggregate (BoW, **VLAD**, FV)

f(I)

Make it trainable end-to-end



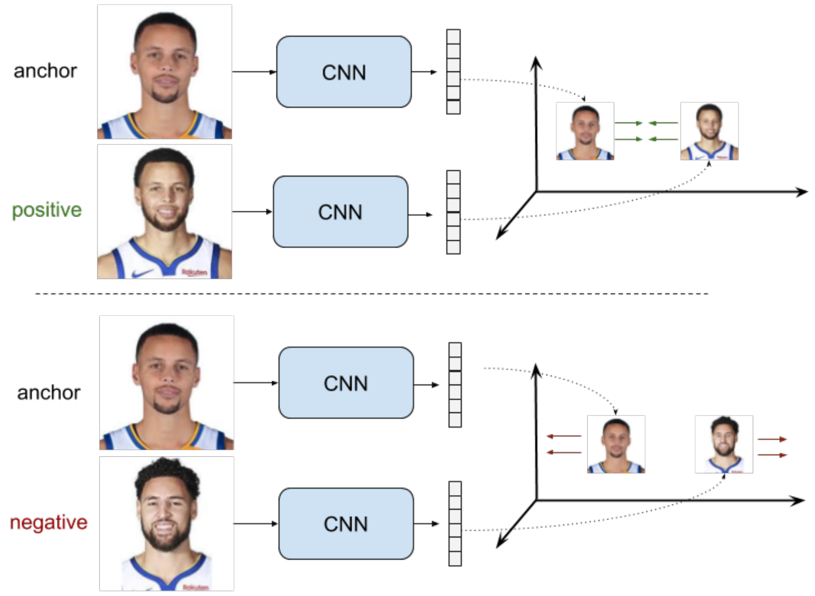
Convolutional layers from AlexNet or VGGNet

Trainable pooling layer

NetVLAD: CNN architecture for weakly supervised place recognition, R Arandjelovic, P Gronat, A Torii, T Pajdla, J Sivic, CVPR 2016

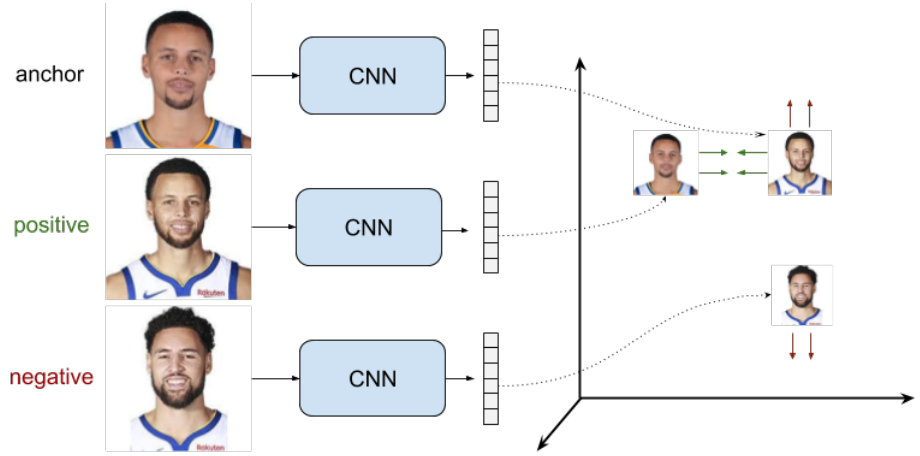
Ranking Loss functions

Pairwise Ranking Loss using Siamese Networks



$$L = \begin{cases} d(r_a, r_p) & if \quad PositivePair \\ max(0, m - d(r_a, r_n)) & if \quad NegativePair \end{cases}$$

Triplet Ranking Loss



$$L(r_a, r_p, r_n) = max(0, m + d(r_a, r_p) - d(r_a, r_n))$$

Results on Dataset Tokyo 24/7 [Torii et al. 15]

Day query

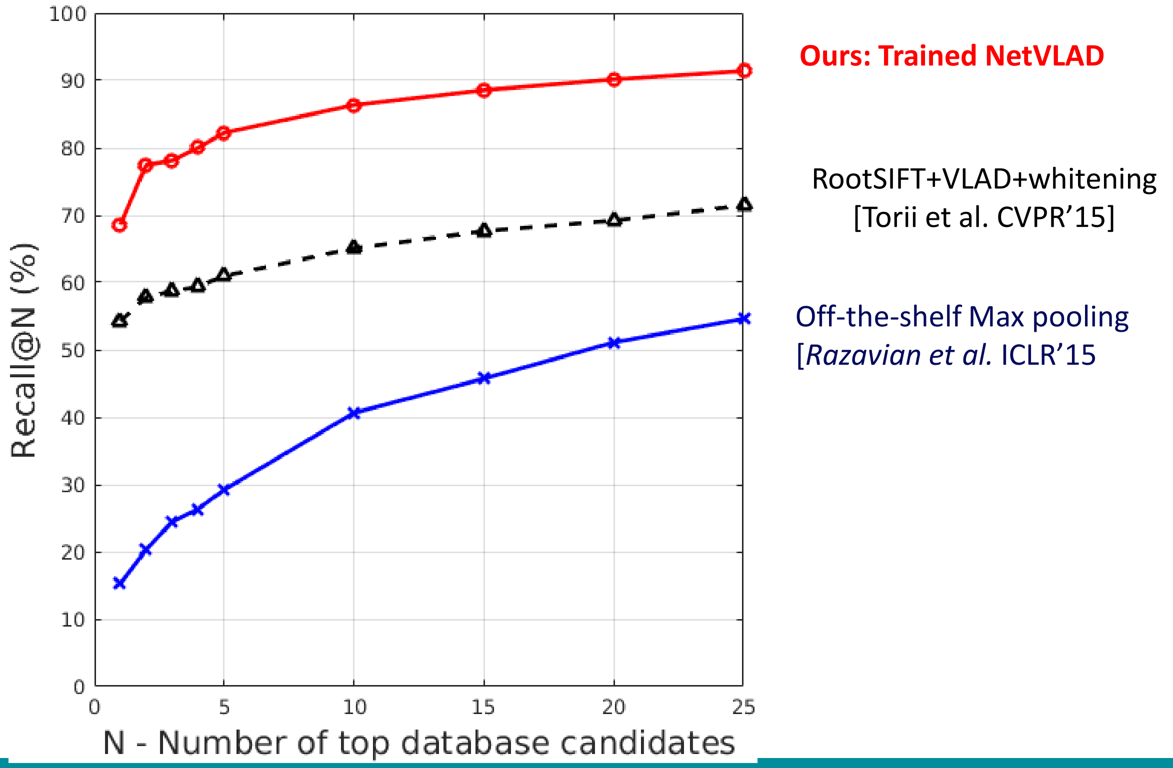
Sunset query

Night query

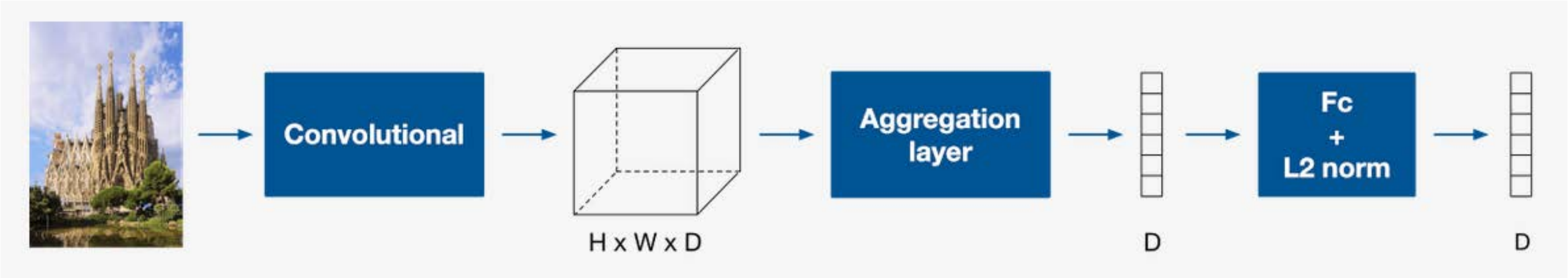
Database image

Database: 76k images from Street View
Queries: 315 images from mobile phone cameras

	recall@5
Previous state-of-the-art	60.9%
Trained NetVLAD	82.2%
Relative improvement	35.0%



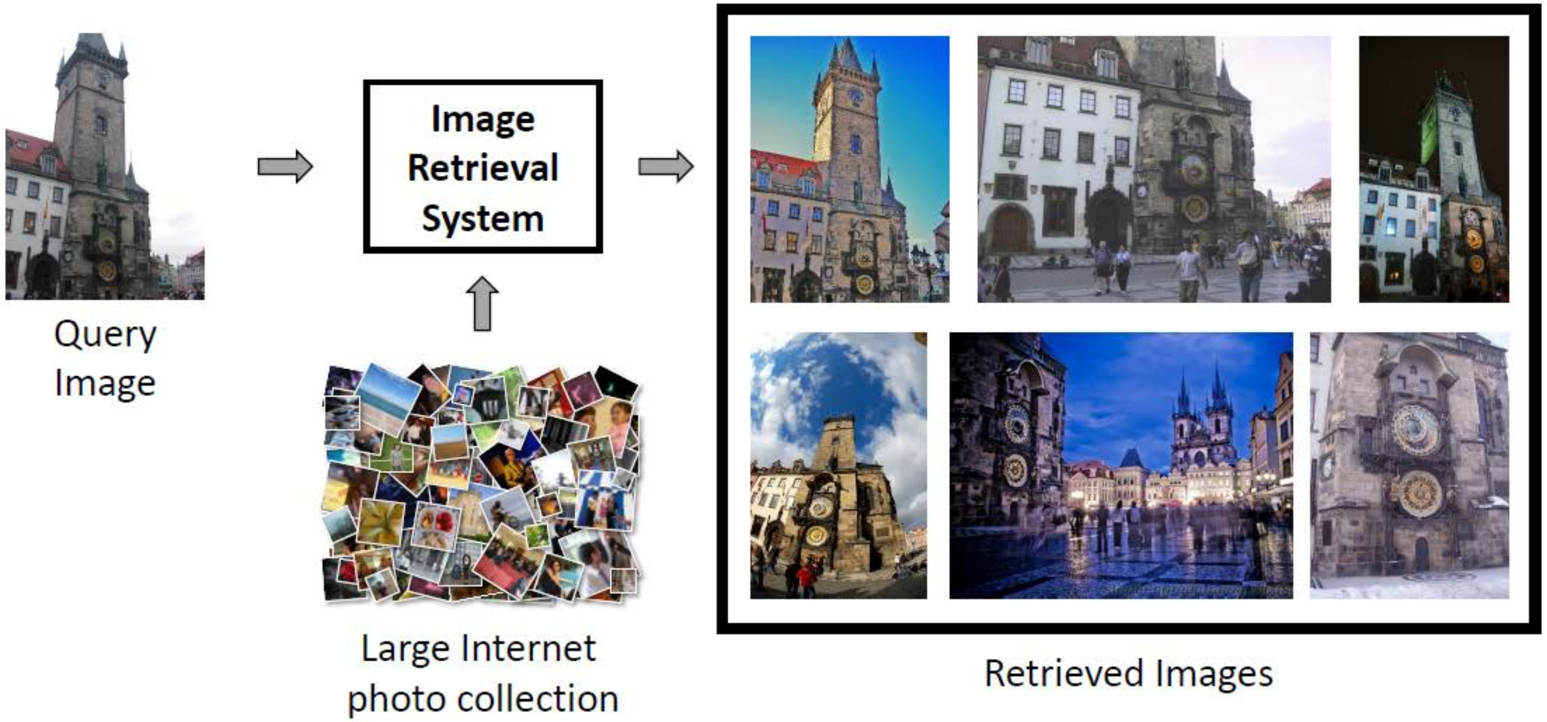
General Deep Learning architecture for coming up with feature vector for global search



5 key choices

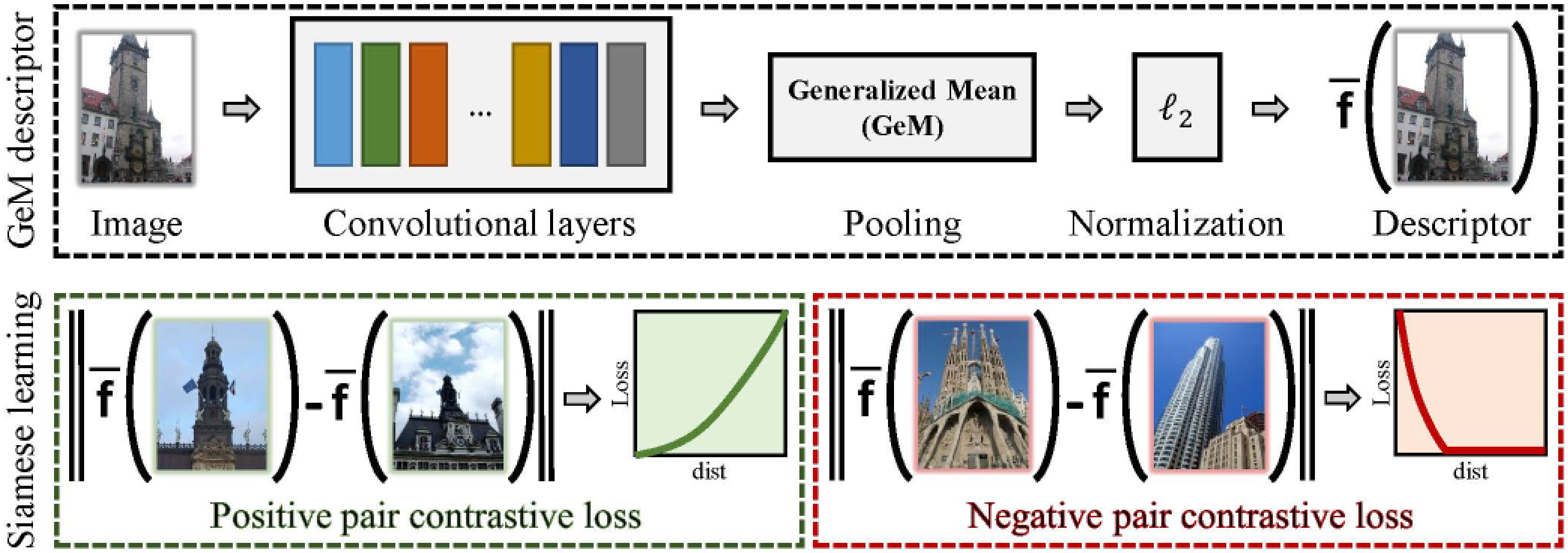
1. Feature Computation
2. Aggregation Layer
3. Normalization
4. Error Metric
5. Training

Visual Retrieval with Compact Image Representations



Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O.,



Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Pooling for Image Representation

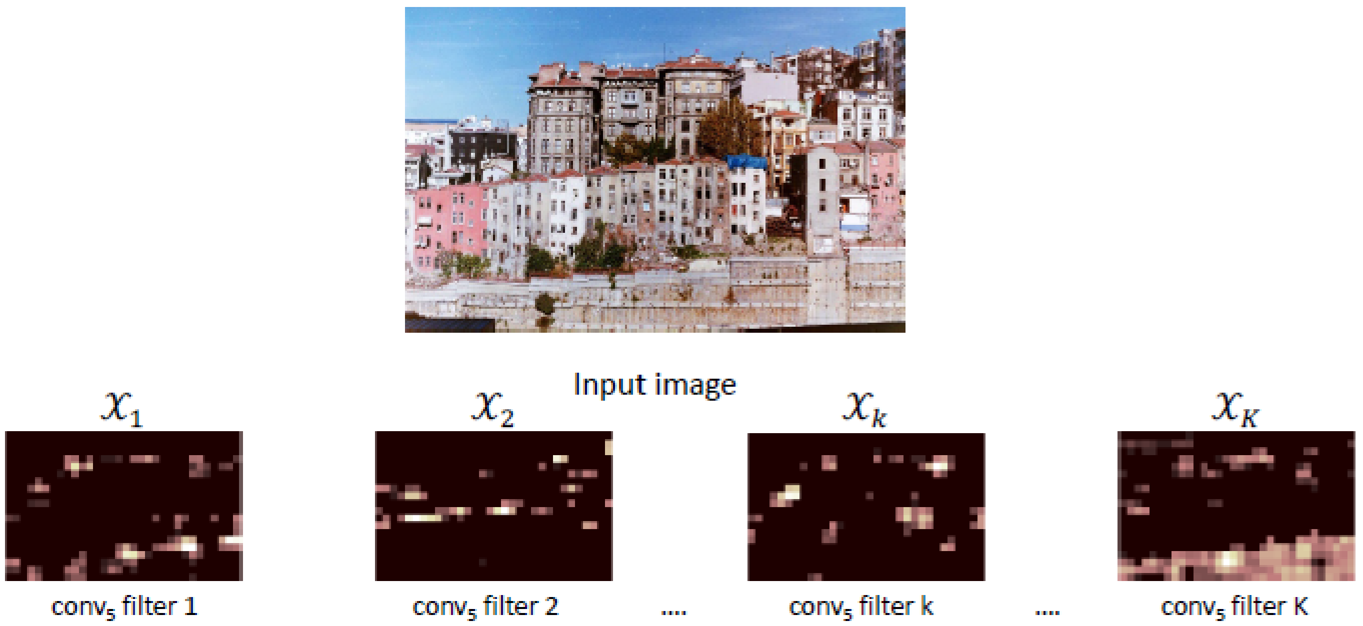


Image descriptor: $\boldsymbol{f} = [f_1 \dots f_k \dots f_K]$

Max pooling (MAC): $f_k = \max_{x \in \mathcal{X}_k} x$

Sum pooling (SpOC): $f_k = \frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x$

Generalized-mean pooling (GeM):

$$f_k = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^p \right)^{\frac{1}{p}} \quad \begin{array}{l} p \to \infty \text{ MAC} \\ p = 1 \text{ SPoC} \end{array}$$

- Observation that max-pooling is more invariant to scale changes,
- Whereas sum-pooling is less sensitive to distractors in the feature maps
- Max-pooling and sum pooling are fixed.
- Generalized mean pooling learns the parameter p and out-performs both Max pooling and Sum-pooling

Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Generalized Mean Pooling (GeM)

TABLE 1
Performance (mAP) comparison after CNN fine-tuning for different pooling layers. GeM is evaluated with a single shared pooling parameter or multiple pooling parameters (one for each feature map), which are either fixed or learned. A single value or a range is reported in the case of a single or multiple parameters, respectively. Results reported with AlexNet and with the use of L_w. The best performance highlighted in **bold**.

Pooling	Initial p	Learned p	Oxford5k	Oxford105k	Paris6k	Paris106k	Holidays	Hol101k
MAC	inf	–	62.2	52.8	68.9	54.7	78.4	66.0
SPoC	1	–	61.2	54.9	70.8	58.0	79.9	70.6
GeM	3	–	**67.9**	60.2	74.8	61.7	83.2	73.3
	[2, 5]	–	66.8	59.7	74.1	60.8	**84.0**	73.6
	[2, 10]	–	65.6	57.8	72.2	58.9	81.9	71.9
	3	2.32	67.7	**60.6**	**75.5**	**62.6**	83.7	**73.7**
	3	[1.0, 6.5]	66.3	57.8	74.0	60.5	83.2	72.7
	[2, 10]	[1.6, 9.9]	65.3	56.4	71.4	58.6	81.4	70.8

Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Feature Vector Dimensionality Reduction using PCA's

High dimensional sparse
BOW image representation

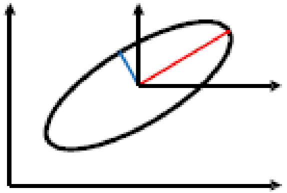
*Search is done using inverted files

PCA

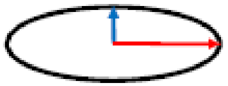
128 dimensional dense
image representation

*Search is done using (approximate) nearest-neighbors

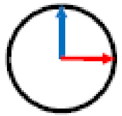
- Centering – emphasize negative evidence, higher importance of jointly missing visual words



- PCA rotation – decorrelating and allowing to remove least informative dimensions

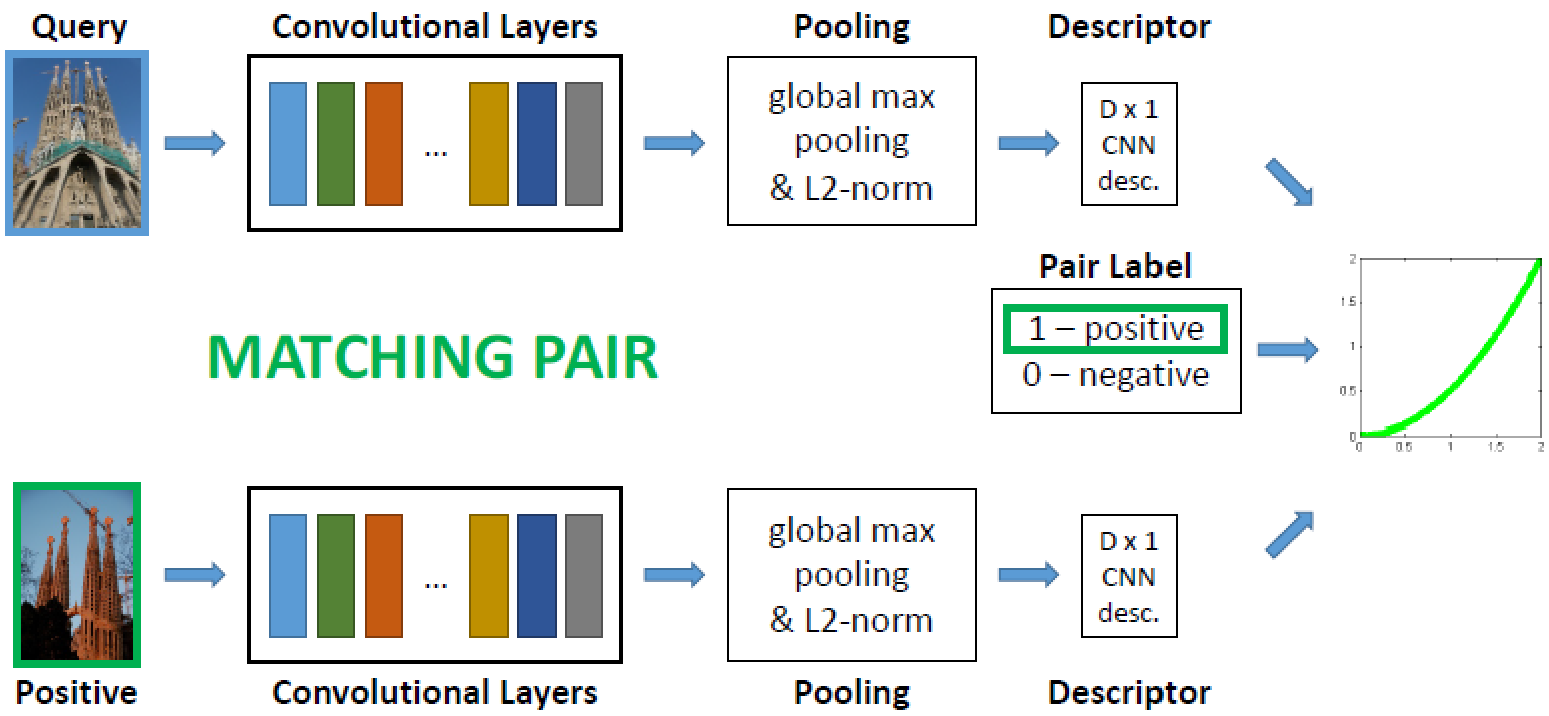


- Whitening – addresses over-counting (burstiness, co-occurence)



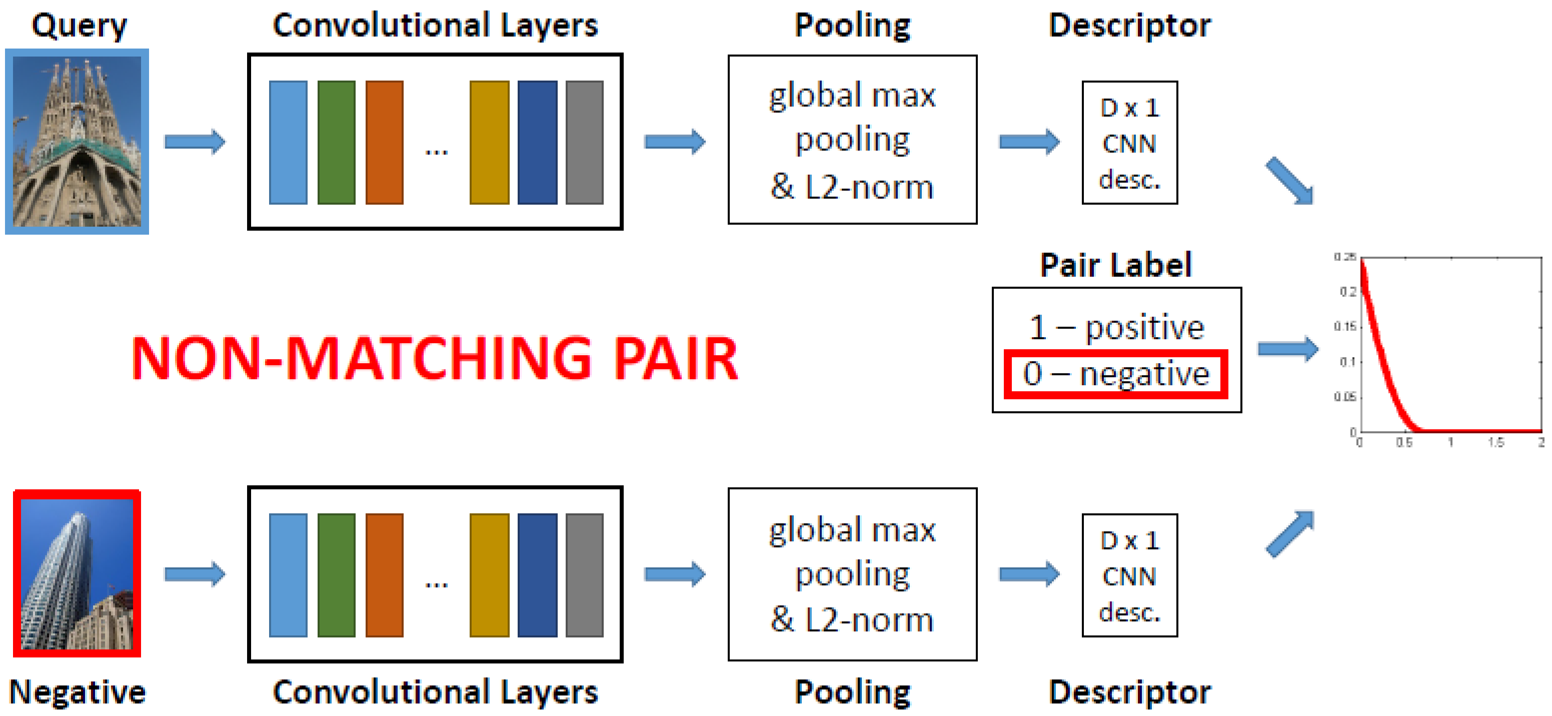
Jegou, Chum: Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening, ECCV 2012

CNN Siamese Learning



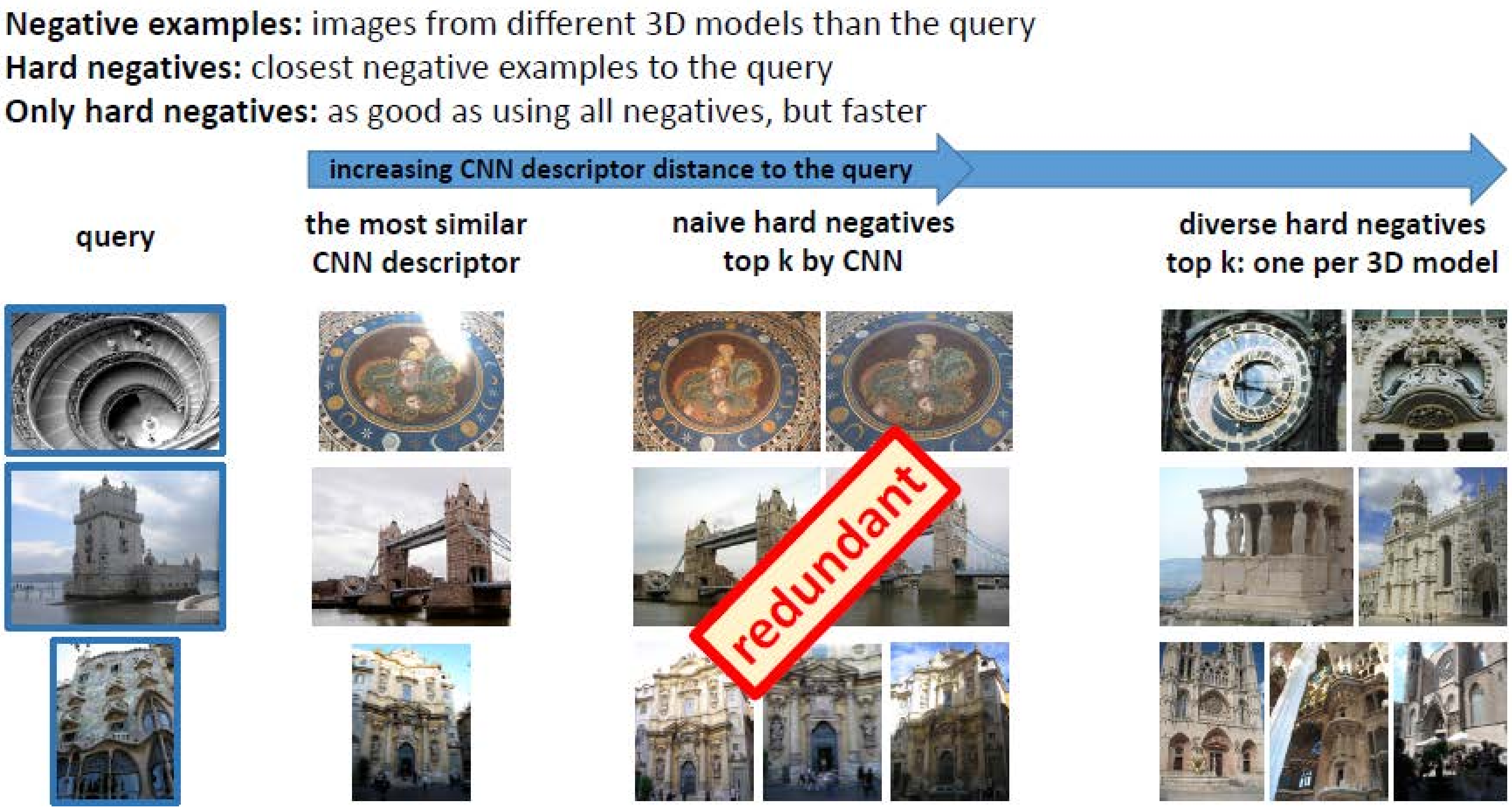
Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

CNN Siamese Learning



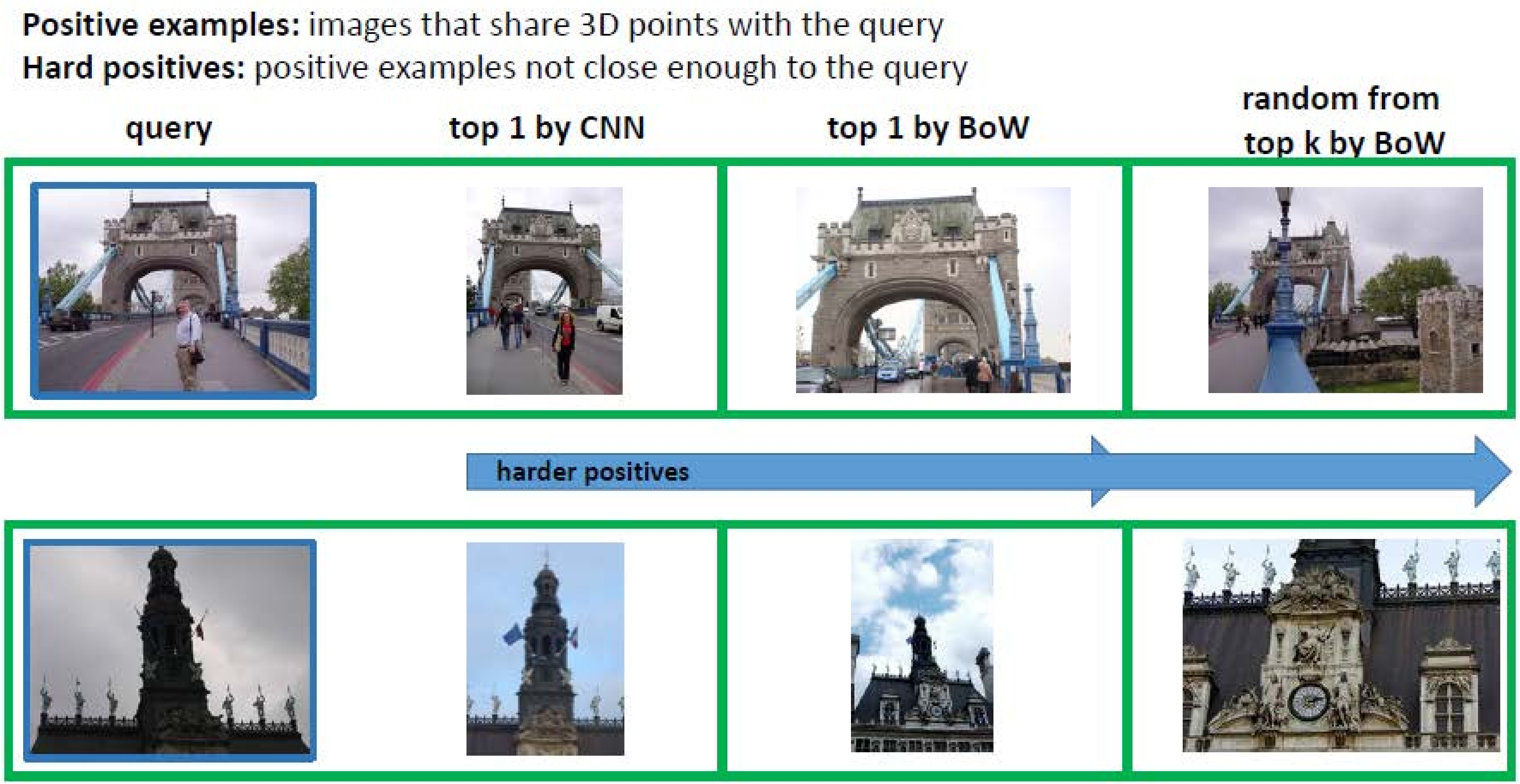
Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Choosing hard negative samples



Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Choosing hard positive samples



Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Training performance comparison based on positive and negative example selection

Choosing harder positive and negatives improves training considerably:

- Harder positives which are not close to the query
- Harder negatives which are closest to positives

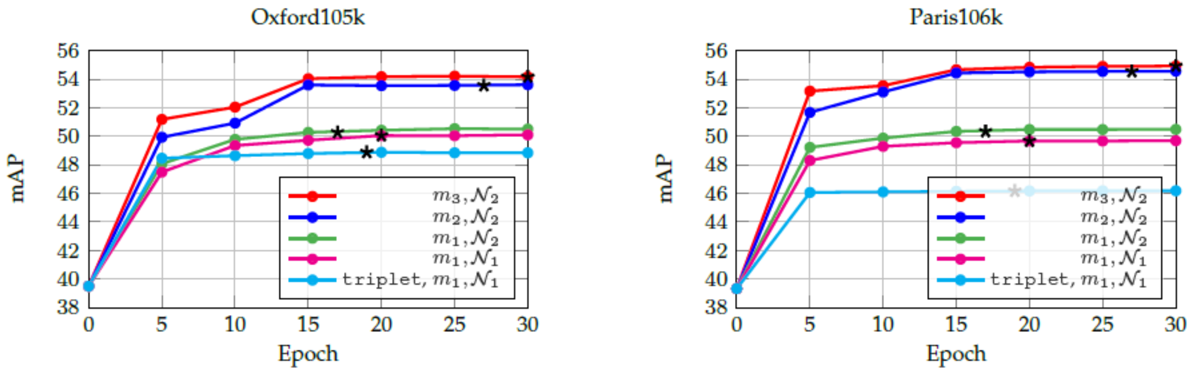


Fig. 7. Performance comparison of methods for positive and negative example selection. Evaluation is performed with AlexNet MAC on Oxford105k and Paris106k datasets. The plot shows the evolution of mAP with the number of training epochs. Epoch 0 corresponds to the off-the-shelf network. All approaches use the contrastive loss, except if otherwise stated. The network with the best performance on the validation set is marked with ⋆.

Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

Fine tuning CNN Image Retrieval – overall results

TABLE 5
Performance (mAP) comparison with the state-of-the-art image retrieval using VGG and ResNet (Res) deep networks, and using local features.
F-tuned: Use of the fine-tuned network (yes), or the off-the-shelf network (no), not applicable for the methods using local features (n/a).
Dim: Dimensionality of the final compact image representation, not applicable (n/a) for the BoW based methods due to their sparse representation.
Our methods are marked with ⋆ and they are always accompanied by the multi-scale representation and our learned whitening L_w.
Previous state of the art is highlighted in **bold**, new state of the art in **red outline**. Best viewed in color.

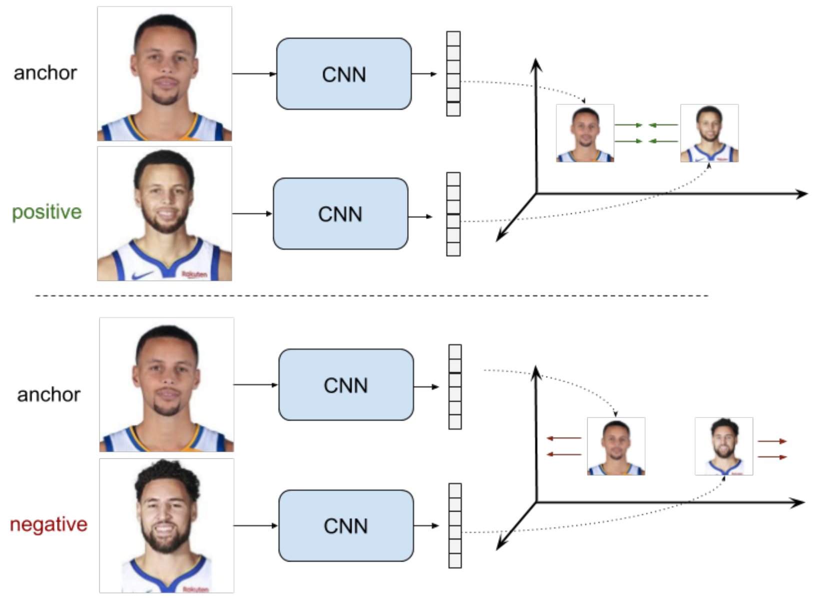
Net	Method	F-tuned	Dim	Oxford5k	Oxford105k	Paris6k	Paris106k	Holidays	Hol101k
Compact representation using deep networks									
VGG	MAC [9]†	no	512	56.4	47.8	72.3	58.0	79.0	66.1
	SPoC [10]†	no	512	68.1	61.1	78.2	68.4	83.9	**75.1**
	CroW [11]	no	512	70.8	65.3	79.7	72.2	85.1	–
	R-MAC [12]	no	512	66.9	61.6	83.0	75.7	86.9‡	–
	BoW-CNN [48]	no	n/a	73.9	59.3	82.0	64.8	–	–
	NetVLAD [16]	no	4096	66.6	–	77.4	–	88.3	–
	NetVLAD [16]	yes	512	67.6	–	74.9	–	86.1	–
	NetVLAD [16]	yes	4096	71.6	–	79.7	–	87.5	–
	Fisher Vector [49]	yes	512	81.5	76.6	82.4	–	–	–
	R-MAC [26]	yes	512	83.1	78.6	87.1	79.7	89.1	–
	⋆ GeM	yes	512	**87.9**	**83.3**	**87.7**	**81.3**	**89.5**	**79.9**
Res	R-MAC [12]‡	no	2048	69.4	63.7	85.2	77.8	91.3	–
	R-MAC [27]	yes	2048	86.1	82.8	**94.5**	**90.6**	**94.8**	–
	⋆ GeM	yes	2048	**87.8**	**84.6**	92.7	86.9	93.9	**87.9**
Re-ranking (R) and query expansion (QE)									
n/a	BoW+R+QE [36]	n/a	n/a	82.7	76.7	80.5	71.0	–	–
	BoW-fVocab+R+QE [59]	n/a	n/a	84.9	79.5	82.4	77.3	75.8	–
	HQE [38]	n/a	n/a	88.0	84.0	82.8	–	–	–
VGG	CroW+QE [11]	no	512	74.9	70.6	84.8	79.4	–	–
	R-MAC+R+QE [12]	no	512	77.3	73.2	86.5	79.8	–	–
	BoW-CNN+R+QE [48]	no	n/a	78.8	65.1	84.8	64.1	–	–
	R-MAC+QE [26]	yes	512	89.1	87.3	91.2	86.8	–	–
	⋆ GeM+αQE	yes	512	**91.9**	**89.6**	**91.9**	**87.6**	–	–
Res	R-MAC+QE [12]‡	no	2048	78.9	75.5	89.7	85.3	–	–
	R-MAC+QE [27]	yes	2048	90.6	89.4	**96.0**	**93.2**	–	–
	⋆ GeM+αQE	yes	2048	**91.0**	**89.5**	95.5	91.9	–	–

†: Our evaluation of MAC and SPoC with PCA_w and with the off-the-shelf network.
‡: Evaluation of R-MAC by [27] with the off-the-shelf network.

Fine-tuning CNN Image Retrieval with No Human Annotation,
Radenović F., Tolias G., Chum O., TPAMI 2018 [arXiv]

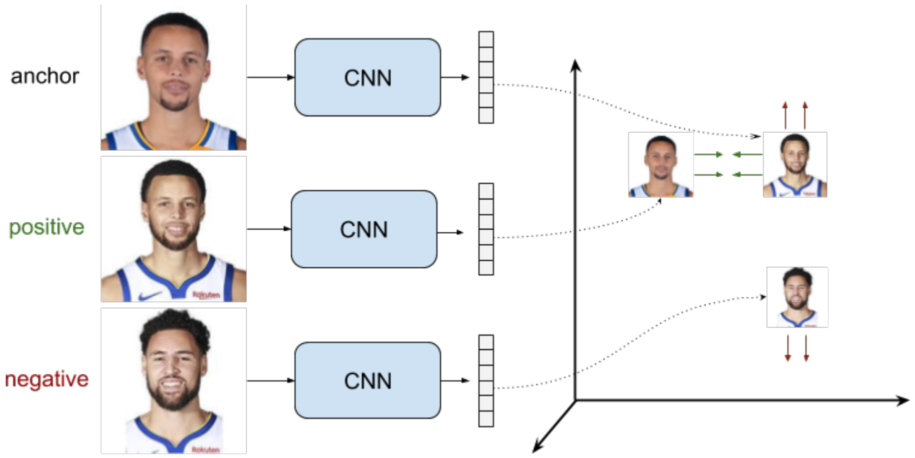
Ranking Loss functions

Pairwise Ranking Loss using Siamese Networks



$$L = \begin{cases} d(r_a, r_p) & if \quad PositivePair \\ max(0, m - d(r_a, r_n)) & if \quad NegativePair \end{cases}$$

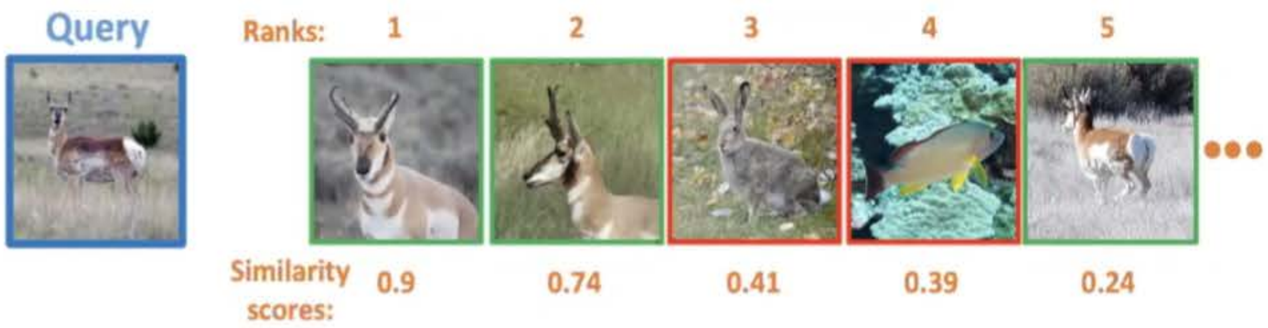
Triplet Ranking Loss



$$L(r_a, r_p, r_n) = max(0, m + d(r_a, r_p) - d(r_a, r_n))$$

Measuring Image Retrieval Performance

- Average Precision used to benchmark retrieval systems
- **Non-differentiable ranking**, so cannot train end-to-end directly
- **Our Goal** – Optimise a smoothed version of the Average Precision Metric

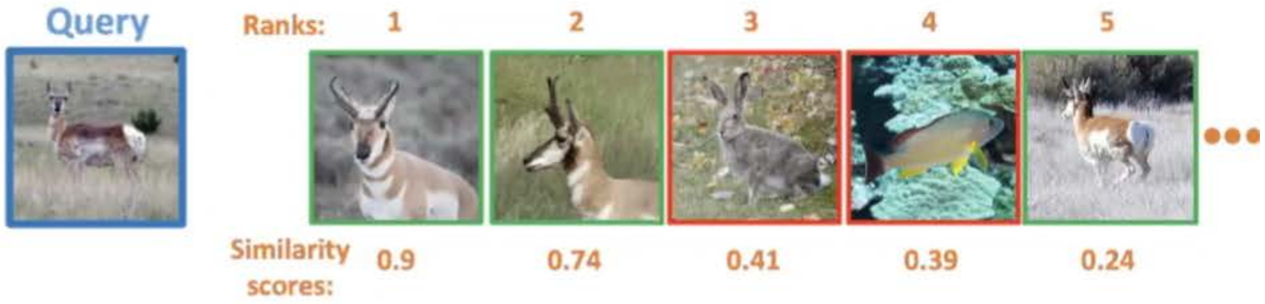


$$AP_q = \frac{1}{|S_p|} \sum_{i \in S_p} \frac{\mathcal{R}^+(i, S_p)}{\mathcal{R}(i, S_\Omega)} \rightarrow \; = \frac{1}{3}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5}\right) \approx 0.87$$

Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020

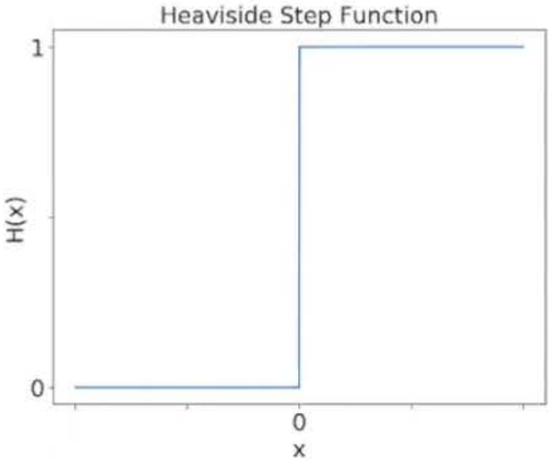
Smoothing the Average Precision Loss (Smooth AP-Loss)

- Non-differentiable ranking
- Find the rank of the first number



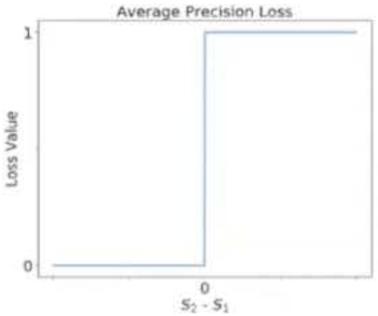
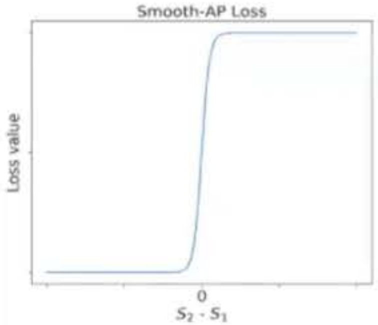
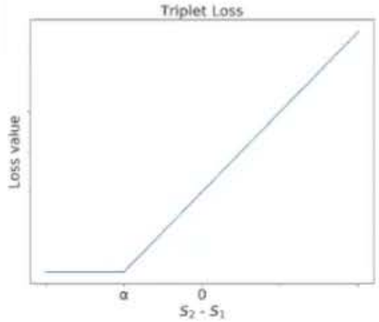
	$\xrightarrow{-0.24}$	$\xrightarrow{H(x)}$	$\xrightarrow{\text{sum}}$
0.24	0	1	5
0.74	0.50	1	
0.39	0.15	1	
0.41	0.17	1	
0.90	0.56	1	

$$H(x) = \begin{cases} 0 & if\ x < 0 \\ 1 & if\ x \geq 0 \end{cases}$$



Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020

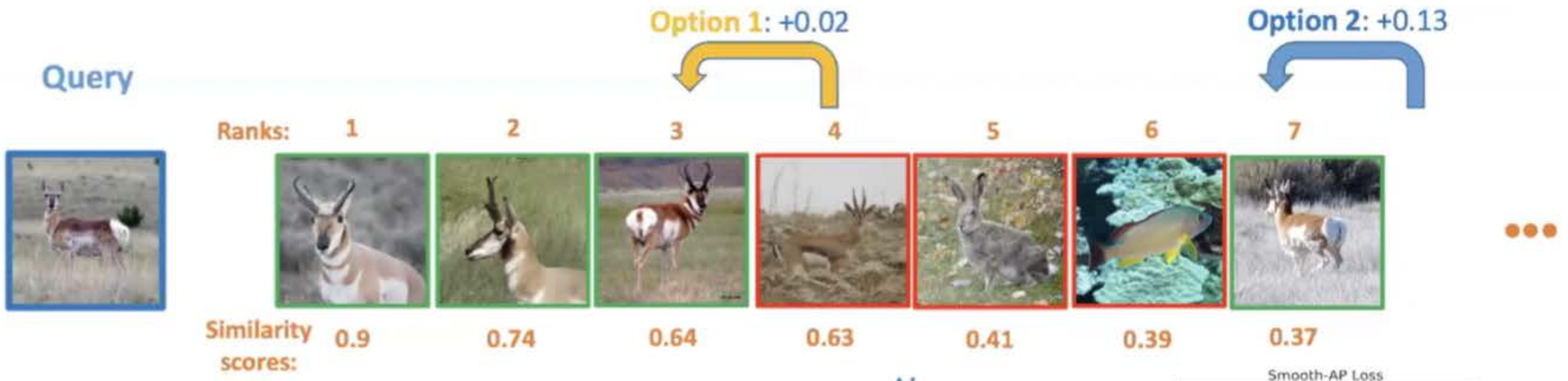
Smooth –AP Loss

			Differentiable?	Optimises Ranking metric?
Average Precision Loss $\mathcal{L}_{AP} = (1 - AP_q)$		$\mathcal{L}_{AP} \propto H(x) = \begin{cases} 0 & if\ x < 0 \\ 1 & if\ x \geq 0 \end{cases}$	✗	✓
Smooth-AP Loss		$\mathcal{L}_{Smooth-AP} \propto G(x) = \dfrac{1}{1 + e^{\frac{-x}{\tau}}}$	✓	✓
Ranking Surrogate loss (e.g. Triplet Loss)		$\mathcal{L}_{triplet} \propto \max(S_2 - S_1 + \alpha, 0)$	✓	✗

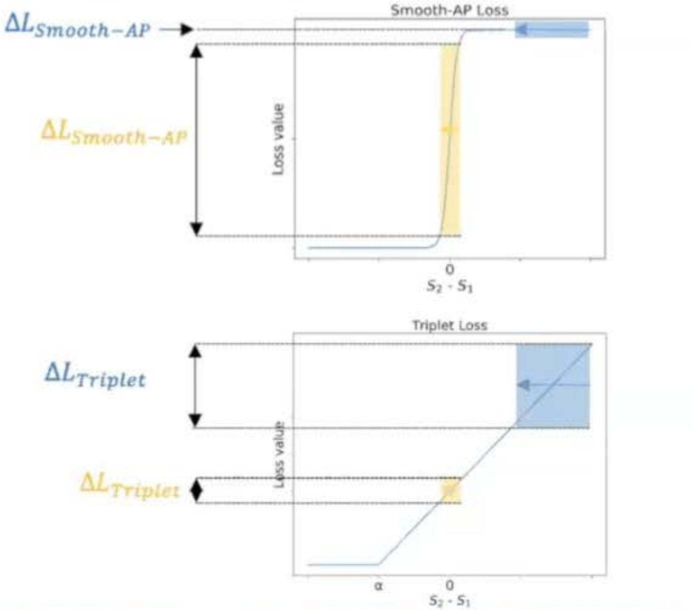
S_2 = positive instance relevance score S_1 = negative instance relevance score

Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020

Smooth –AP Loss



- Smooth-AP optimises a ranking metric
  - **Option 1**: small score change (+0.02), rank change →$\Delta AP > 0$
  - **Option 2**: large score change (+0.13), no rank change →$\Delta AP = 0$
- **Smooth-AP favours the rank change** → larger reduction in loss
- Triplet favours the large score change → larger reduction in loss



S_2 = positive instance relevance score S_1 = negative instance relevance score

Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020

Experiments on INaturalist

- INaturalist
  - ➢ Smooth-AP outperforms previous AP-approximating approaches

	INaturalist			
Recall@K	1	4	16	32
Triplet Semi-Hard (NeurIPS '06)	58.1	75.5	86.8	90.7
Proxy NCA (CVPR '17)	61.6	77.4	87.0	90.6
* FastAP (CVPR '19)	60.6	77.0	87.2	90.6
* Blackbox AP (CVPR '20)	62.9	79.0	88.9	92.1
Smooth-AP BS=224	65.9	80.9	89.8	92.7
Smooth-AP BS=384	**67.2**	**81.8**	**90.3**	**93.1**

* Recent AP-approximating approaches

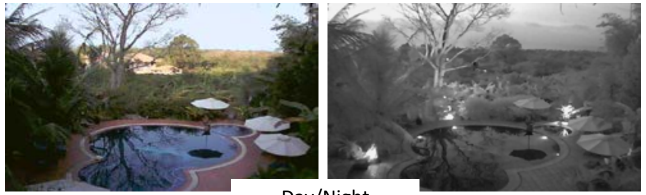
Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval, A. Brown, W. Xie, V. Kalogeiton, A. Zisserman, European Conference on Computer Vision, 2020

SRI International

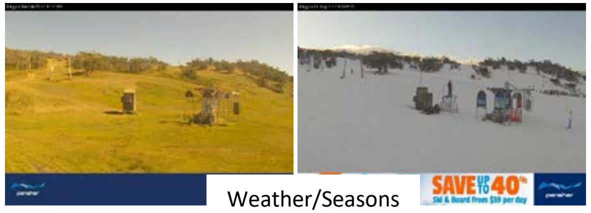
Cross-time Geo-Localization using Semantic Cues

Image-to-Image Visual Localization using semantic cues with attention

- **Goal:** Geo-tag a position for a given monocular **query image** by retrieval from a database of images of **known locations,** in the presence of large appearance changes across weather and illumination variations.


Day/Night

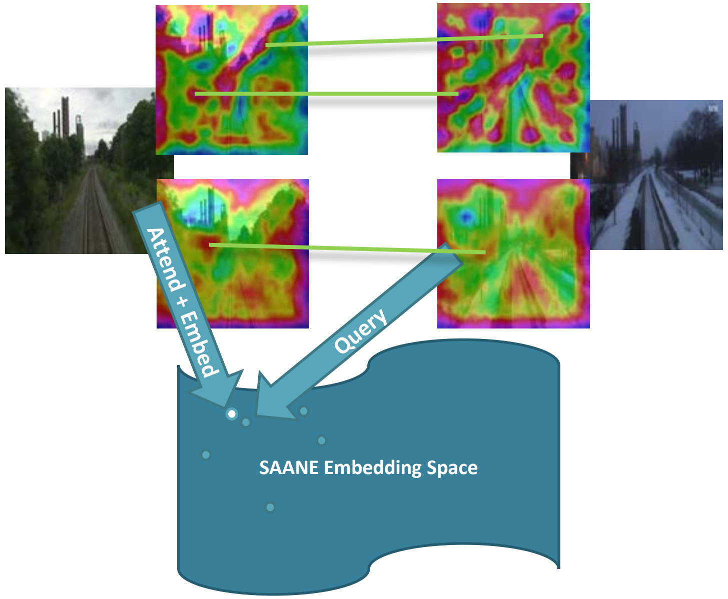
Day/Night

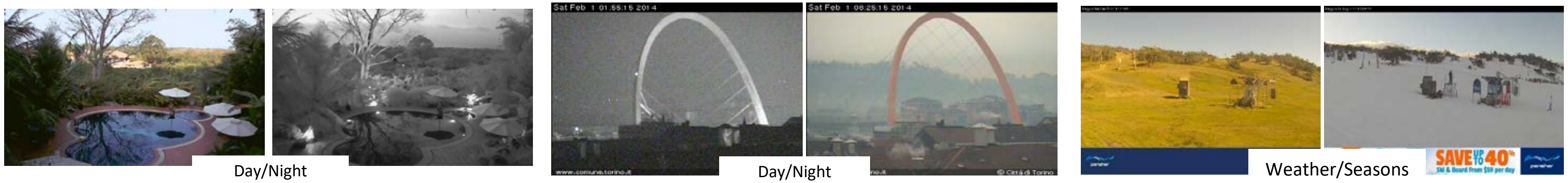


Weather/Seasons

Image-to-Image Visual Localization

- We propose to use embedding for this problem: A deep-learned compact Euclidean space where distances directly correspond to a measure of data similarity.
- Training data: ~2 million images collected from 2,685 static webcams.

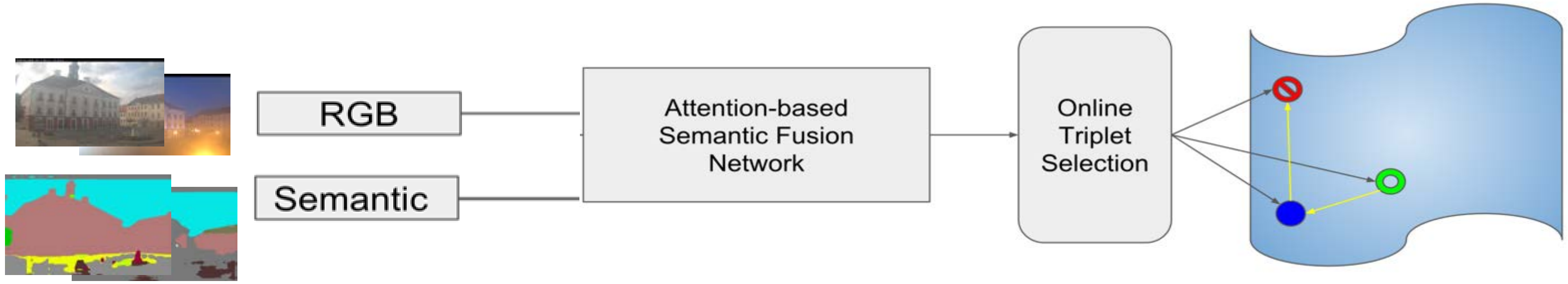




Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

Innovation: Attention-based Semantic-Aware Embedding

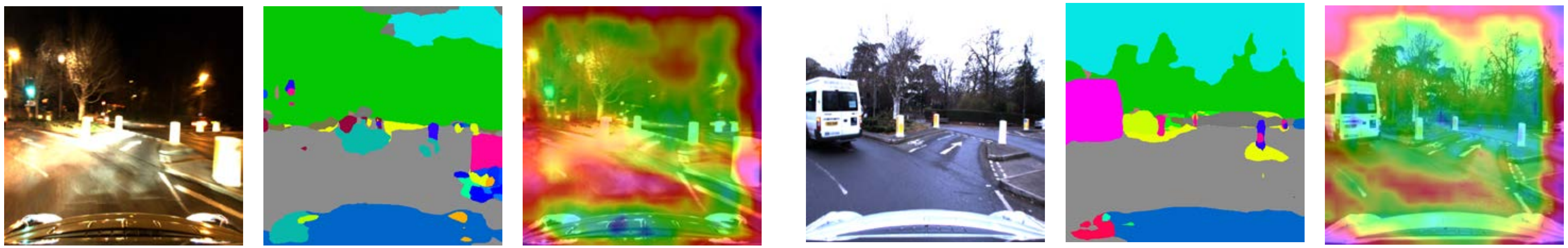
- **Semantic-Aware**: The model incorporates pixel-wise semantic features in learning the image embeddings.
- **Attention-Based**: We train self-attention modules to encourage the model to focus on semantically meaningful spatial regions.
- We evaluate two ways to train attention module: (1) individual attention: on RGB and semantic cue separately, (2) combined attention: on fused feature maps from RGB and semantic cue.



Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

Innovation: Semantically-Aware

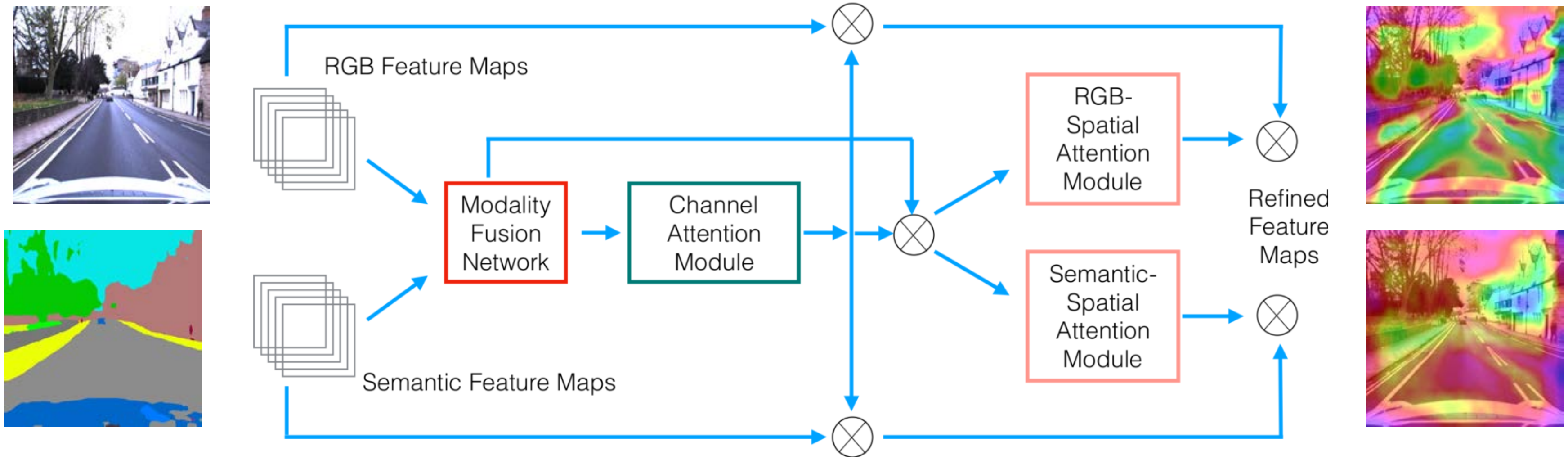
- The model incorporates pixel-wise semantics in learning the image embeddings.
  - Compared to low-level appearance descriptors, the spatial layout of semantic classes in the image yields scene descriptions that have a higher invariance to large changes in viewing conditions, to improve visual localization.
- Ours is the first such model to integrate appearance and semantic features in an end-to-end, principled manner.
- Our results show an 8–15% absolute improvement over our baseline from adding semantic information.



Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

Innovation: Attention-Based

- We train a novel formulation of the Convolutional Block Attention Module to encourage our model to focus on semantically-consistent spatial regions.
- The attention network operates in two steps:
  - First, a single attention map is computed for the fused (appearance + semantic) features across the channel dimension to due an initial, multimodal refinement.
  - Second, separate appearance and semantic spatial attention maps are computed to produce the final, refined feature maps.
- Our attention module combined with semantics gives an additional 4% absolute improvement on average.



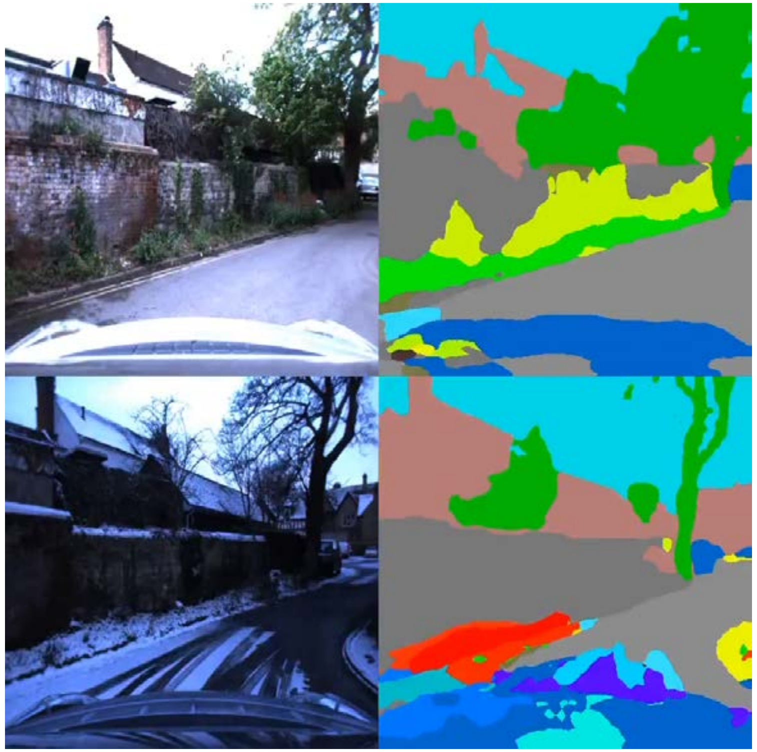
Sanghyun Woo, et al. "CBAM: Convolutional block attention module." *ECCV*. 2018.

Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

Semantic Segmentation of video over seasons and time

Summer
Video

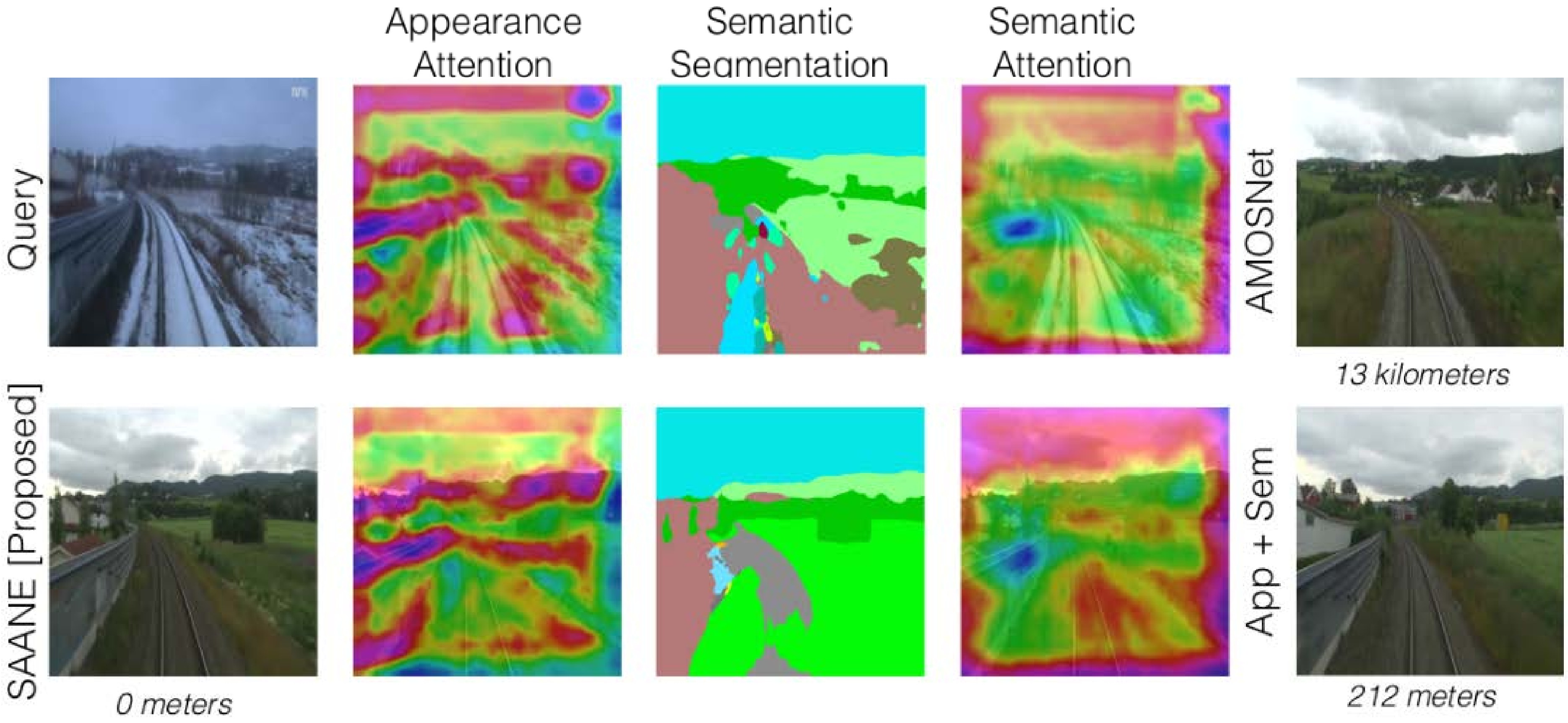
Winter
Video



Semantic
Segmentation
videos

Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

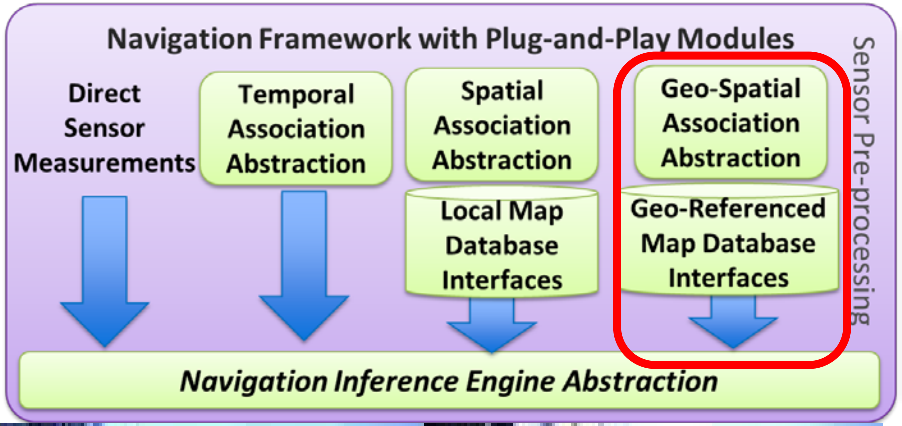
Retrieval: Nordland Summer->Winter



Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization
https://arxiv.org/abs/1812.03402

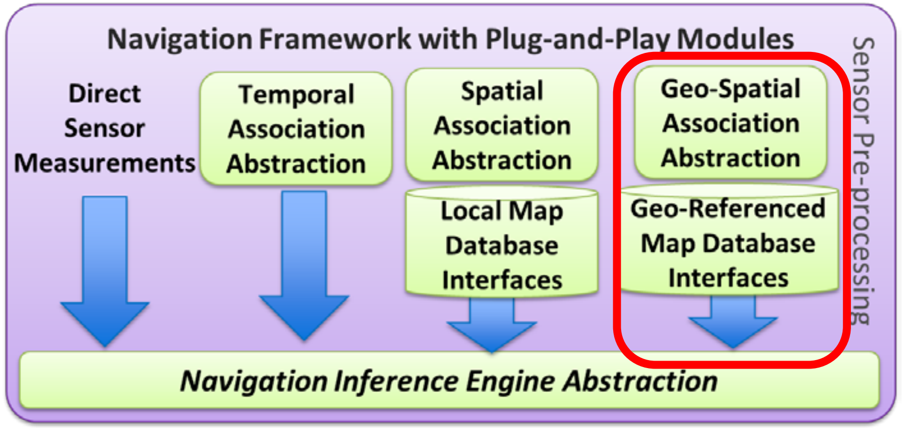
Geo-Spatial Association - Across Seasons

- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.



Geo-Spatial Association - Day & Night

- 2km database, Accuracy can be further improved by position prior, sequential verification, and 2D-3D refinements.



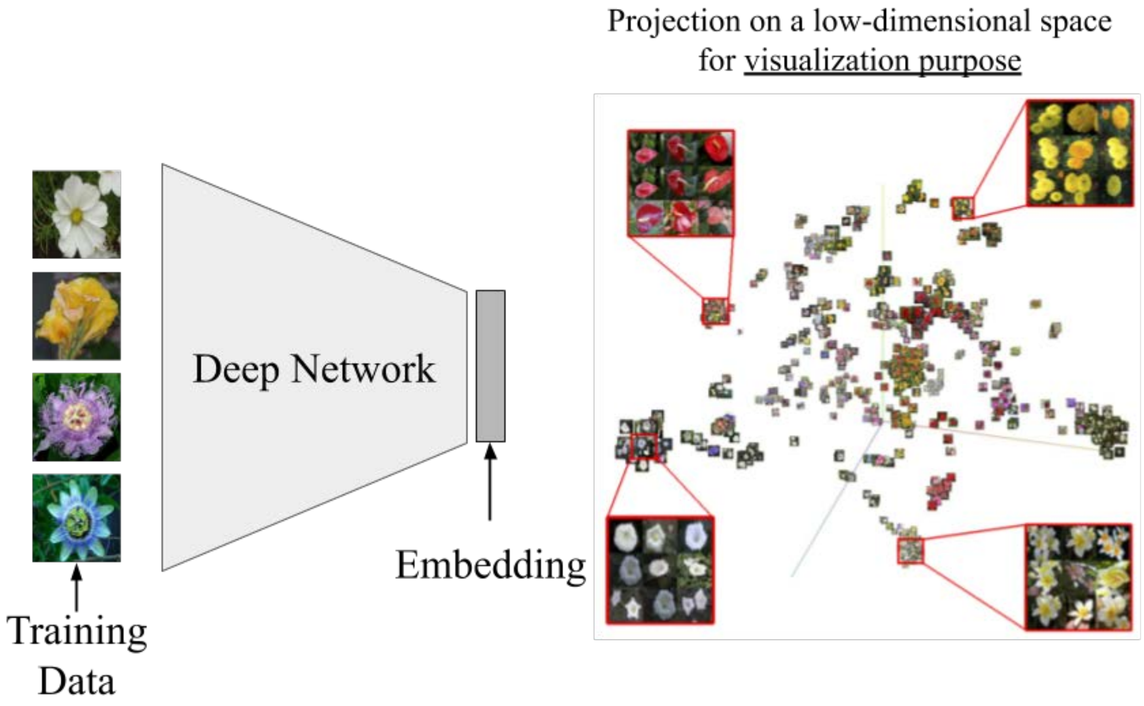
Fine Geo-localization and end-to-end solutions

Cross-season / Cross-time Fine Alignment: Estimating camera pose (Location and Orientation)

Offline: Build image database and 3D Model/ Map for area of regard

Live: Estimate camera pose for a query image:

- **Image retrieval by search from database**
- Detect, describe, and match features to establish correspondences
- Estimate camera pose by resection using 2D-3D (image to model) correspondences



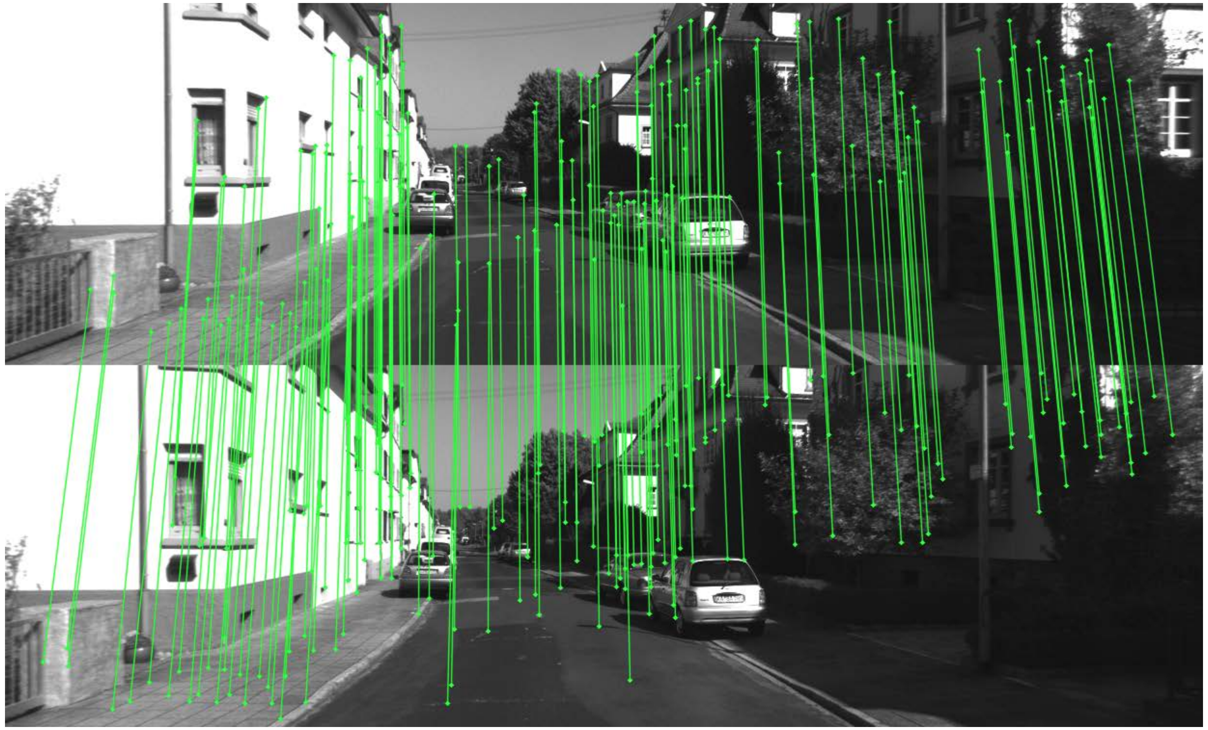
- Jégou, et al. "Aggregating local descriptors into a compact image representation." CVPR, 2010.
- Torii, et al. "24/7 place recognition by view synthesis." CVPR, 2015.
- Mithun, et al. "RGB2LIDAR: Towards Solving Large-Scale Cross-Modal Visual Localization." ACM Multimedia, 2020.

Cross-season / Cross-time Fine Alignment

Offline: Build image database and 3D Model/ Map for area of regard

Live: Estimate camera pose for a query image:

- Image retrieval by search from database
- **Detect, describe, and match features to establish correspondences**
- Estimate camera pose by resection using 2D-3D (image to model) correspondences



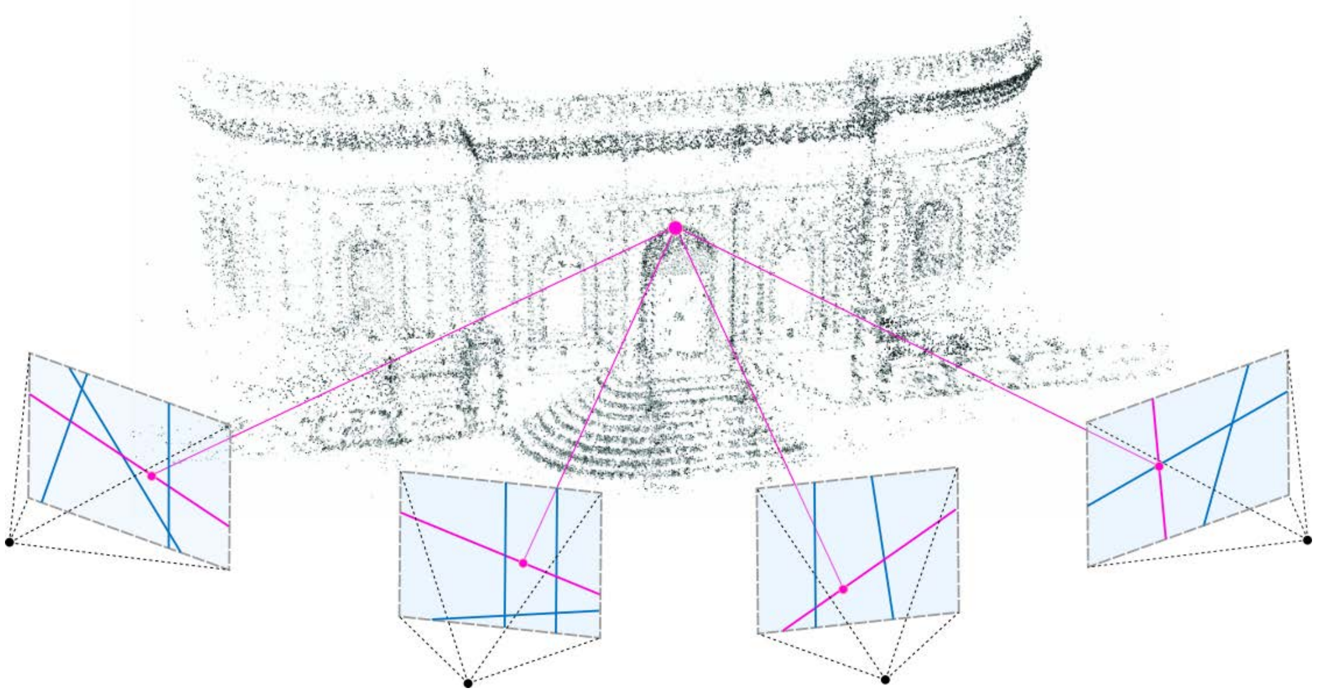
- Lowe. "Distinctive image features from scale-invariant keypoints." IJCV, 2004.
- Rublee, et al. "ORB: An efficient alternative to SIFT or SURF." ICCV, 2011.
- DeTone, et al. "Superpoint: Self-supervised interest point detection and description." CVPR, 2018.
- Sarlin, et al. "Superglue: Learning feature matching with graph neural networks." CVPR, 2020.

Cross-season / Cross-time Fine Alignment

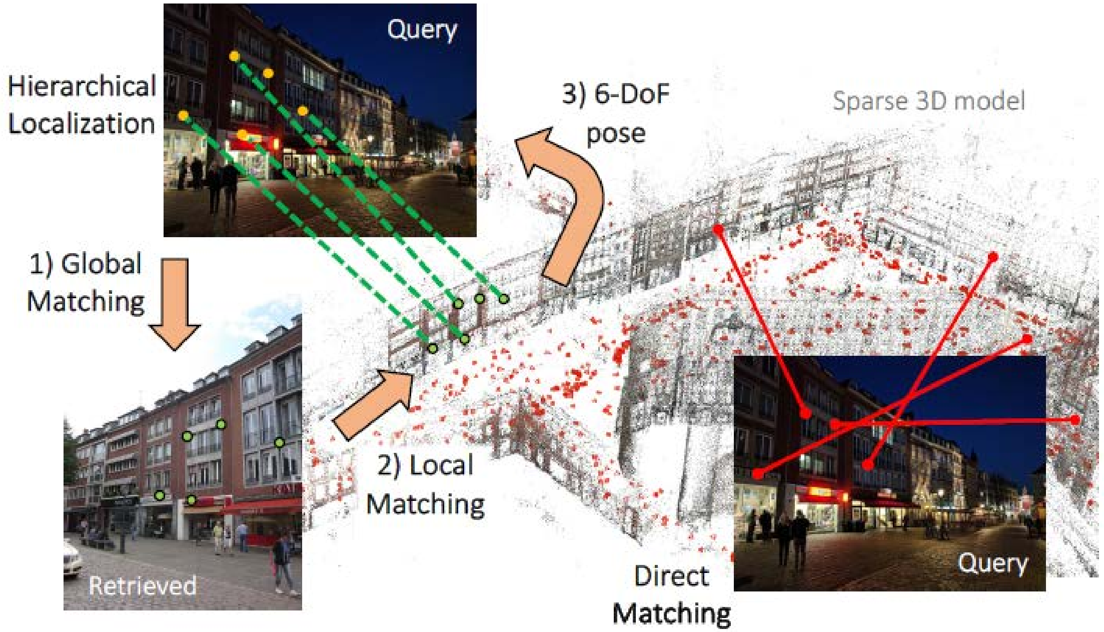
Offline: Build image database and 3D Model/ Map for area of regard

Live: Estimate camera pose for a query image:

- Image retrieval by search from database
- Detect, describe, and match features to establish correspondences
- **Estimate camera pose by resection using 2D-3D (image to model) correspondences**



From Coarse to Fine: Robust Hierarchical Localization at Large Scale

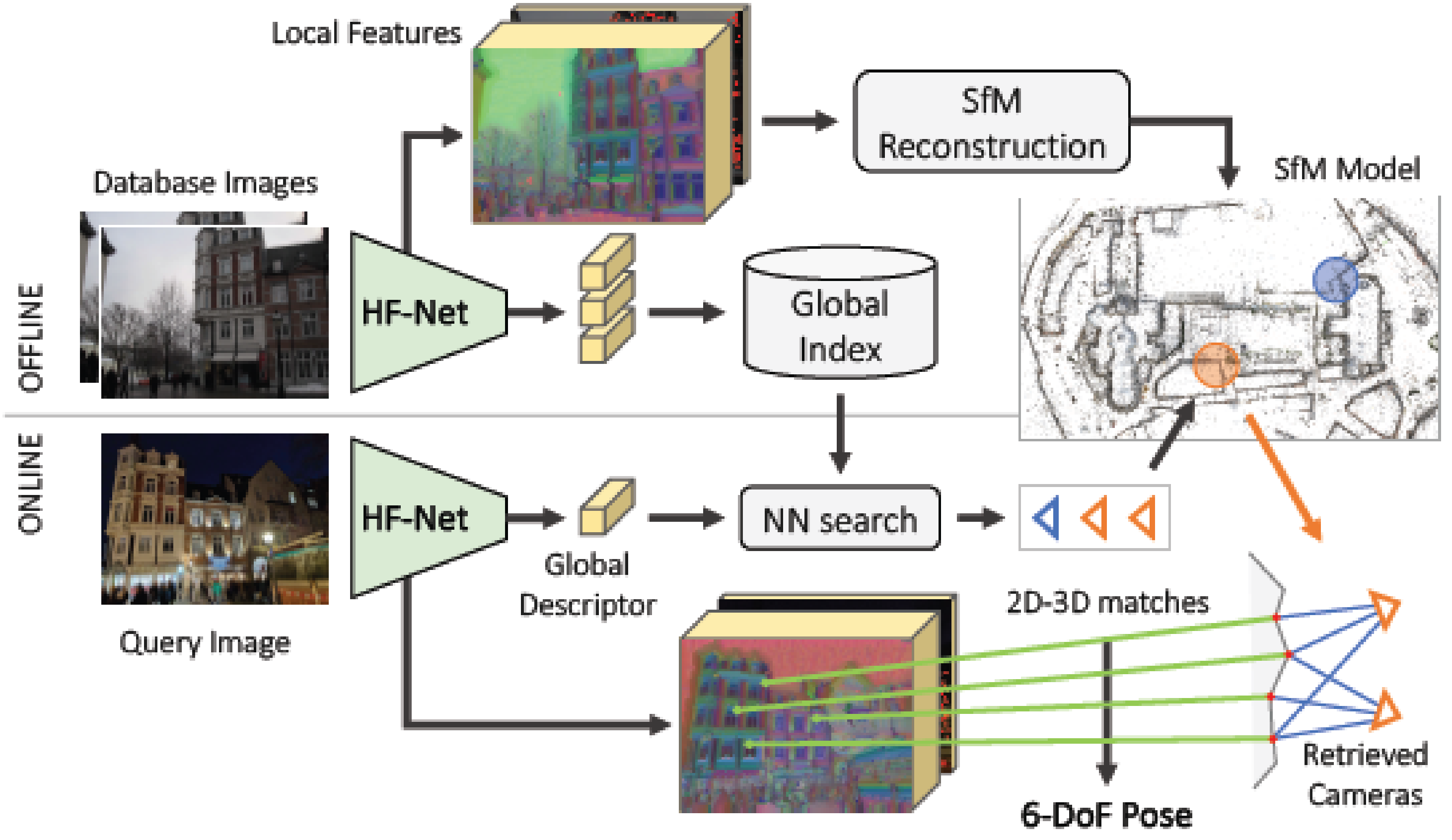


Hierarchical localization:

- A global search first retrieves candidate images,
- Candidate images are subsequently matched using powerful local features to estimate an accurate 6-DoF pose.
- This three-step process is both efficient and robust in challenging situations.

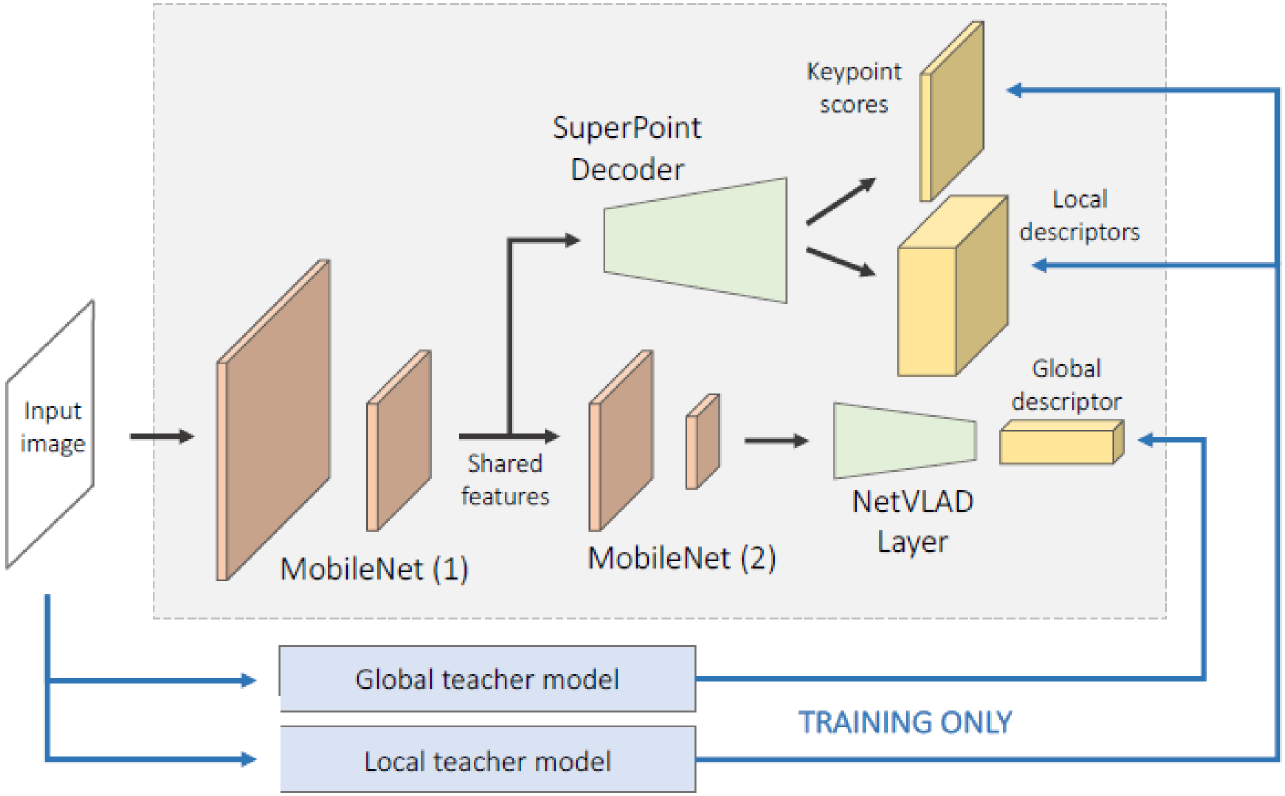
P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, ''From coarse to fine: Robust hierarchical localization at large scale,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, p. 12716.

Hierarchical Localization (using HF-Net)



P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, ''From coarse to fine: Robust hierarchical localization at large scale,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, p. 12716.

Training HF-Net using Teacher models



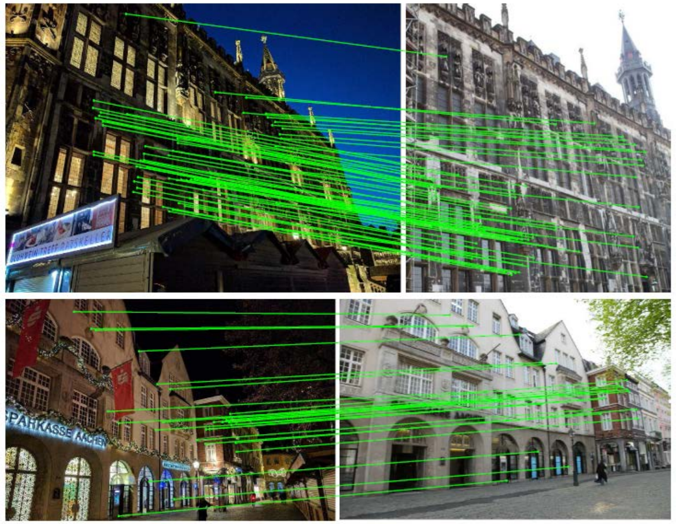
- Local and global descriptors are often trained with metric learning using ground truth positive and negative pairs of local patches and full images.
- These ground truth correspondences are particularly difficult to obtain at the scale required to train large CNNs.
- HF-Net generates three outputs from a single image: a global descriptor, a map of keypoint detection scores, and dense keypoint descriptors.
- All three heads are trained jointly with multi-task distillation from different teacher networks.

P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, ''From coarse to fine: Robust hierarchical localization at large scale,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, p. 12716.

Results

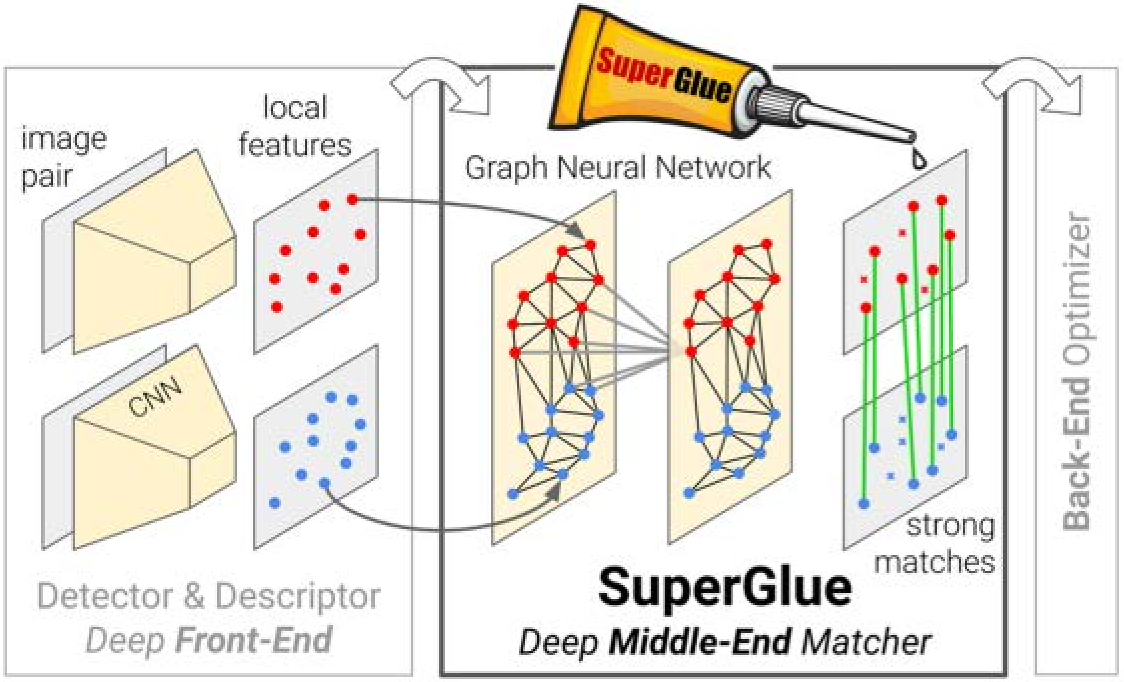
	Aachen		RobotCar				CMU	
	day	night	dusk	sun	night	night-rain	urban	suburban
distance [m]	.25/.50/5.0	0.5/1.0/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0	.25/.50/5.0
orient. [deg]	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10	2/5/10
AS	57.3 / 83.7 / 96.6	19.4 / 30.6 / 43.9	44.7 / 74.6 / 95.9	25.0 / 46.5 / 69.1	0.5 / 1.1 / 3.4	1.4 / 3.0 / 5.2	55.2 / 60.3 / 65.1	20.7 / 25.9 / 29.9
CSL	52.3 / 80.0 / 94.3	24.5 / 33.7 / 49.0	56.6 / 82.7 / 95.9	28.0 / 47.0 / 70.4	0.2 / 0.9 / 5.3	0.9 / 4.3 / 9.1	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1
DenseVLAD	0.0 / 0.1 / 22.8	0.0 / 2.0 / 14.3	10.2 / 38.8 / 94.2	5.7 / 16.3 / 80.2	0.9 / 3.4 / 19.9	1.1 / 5.5 / 25.5	22.2 / 48.7 / 92.8	9.9 / 26.6 / 85.2
NetVLAD	0.0 / 0.2 / 18.9	0.0 / 2.0 / 12.2	7.4 / 29.7 / 92.9	5.7 / 16.5 / 86.7	0.2 / 1.8 / 15.5	0.5 / 2.7 / 16.4	17.4 / 40.3 / 93.2	7.7 / 21.0 / 80.5
SMC	-	-	(53.8 / 83.0 / 97.7)	(46.7 / 74.6 / 95.9)	(6.2 / 18.5 / 44.3)	(8.0 / 26.4 / 46.4)	75.0 / 82.1 / 87.8	44.0 / 53.6 / 63.7
NV+SIFT	82.8 / 88.1 / 93.1	30.6 / 43.9 / 58.2	55.6 / 83.5 / 95.3	46.3 / 67.4 / 90.9	4.1 / 9.1 / 24.4	2.3 / 10.2 / 20.5	63.9 / 71.9 / 92.8	28.7 / 39.0 / 82.1
NV+SP (ours)	79.7 / 88.0 / 93.7	40.8 / 56.1 / 74.5	54.8 / 83.0 / 96.2	51.7 / 73.9 / 92.4	6.6 / 17.1 / 32.2	5.2 / 17.0 / 26.6	91.7 / 94.6 / 97.7	74.6 / 81.6 / 91.4
HF-Net (ours)	75.7 / 84.3 / 90.9	40.8 / 55.1 / 72.4	53.9 / 81.5 / 94.2	48.5 / 69.1 / 85.7	2.7 / 6.6 / 15.8	4.7 / 16.8 / 21.8	90.4 / 93.1 / 96.1	71.8 / 78.2 / 87.1

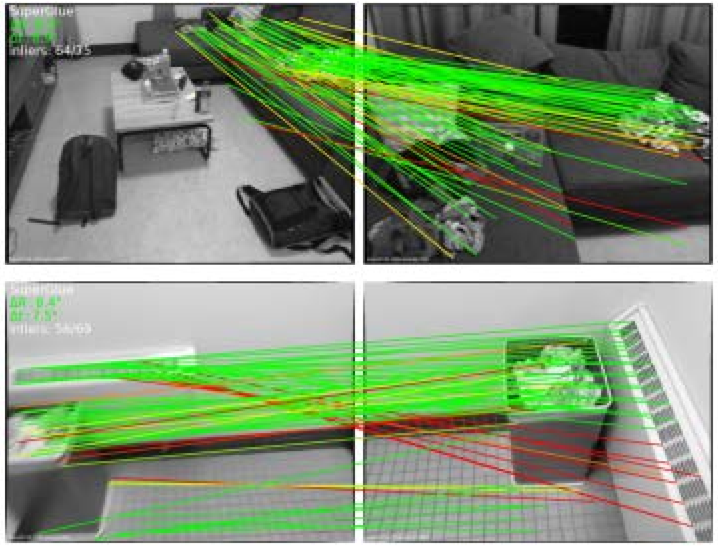
Evaluation of the localization on the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets. Report the recall [%] at different distance and orientation thresholds and highlight for each of them the best and second-best methods. X+Y denotes hierarchical localization with X (Y) as global (local) descriptors. SMC is excluded from the comparison for RobotCar as it uses extra semantic data.



Successful localization with HF-Net on the Aachen Day-Night dataset. Two queries (left) and the retrieved database images with the most inlier matches (right).

SuperGlue = Graph Neural Nets + Matching





- Extreme wide-baseline image pairs in real-time on GPU
- State-of-the-art indoor+outdoor matching with SIFT & SuperPoint

SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

SuperGlue: Fine Geo-localization

A minimal matching pipeline

SuperGlue: context aggregation + matching + filtering

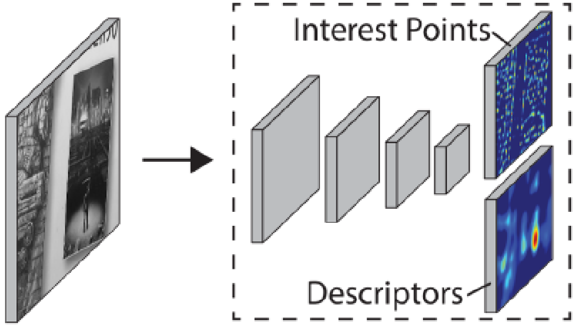
image pair

detection → description → feature matching → outlier filtering → pose estimation

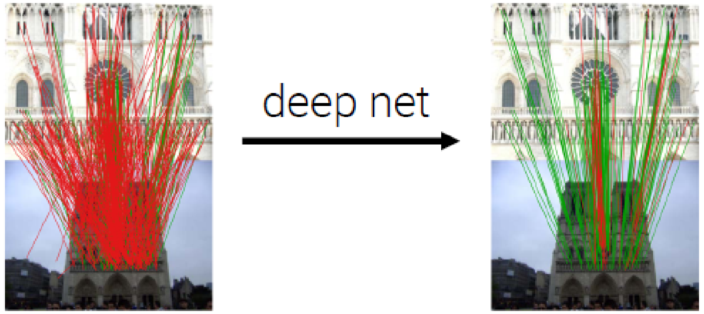
> Classical: SIFT, ORB
> Learned: SuperPoint, D2-Net

Nearest Neighbor Matching

> Heuristics: ratio test, mutual check
> Learned: classifier on set



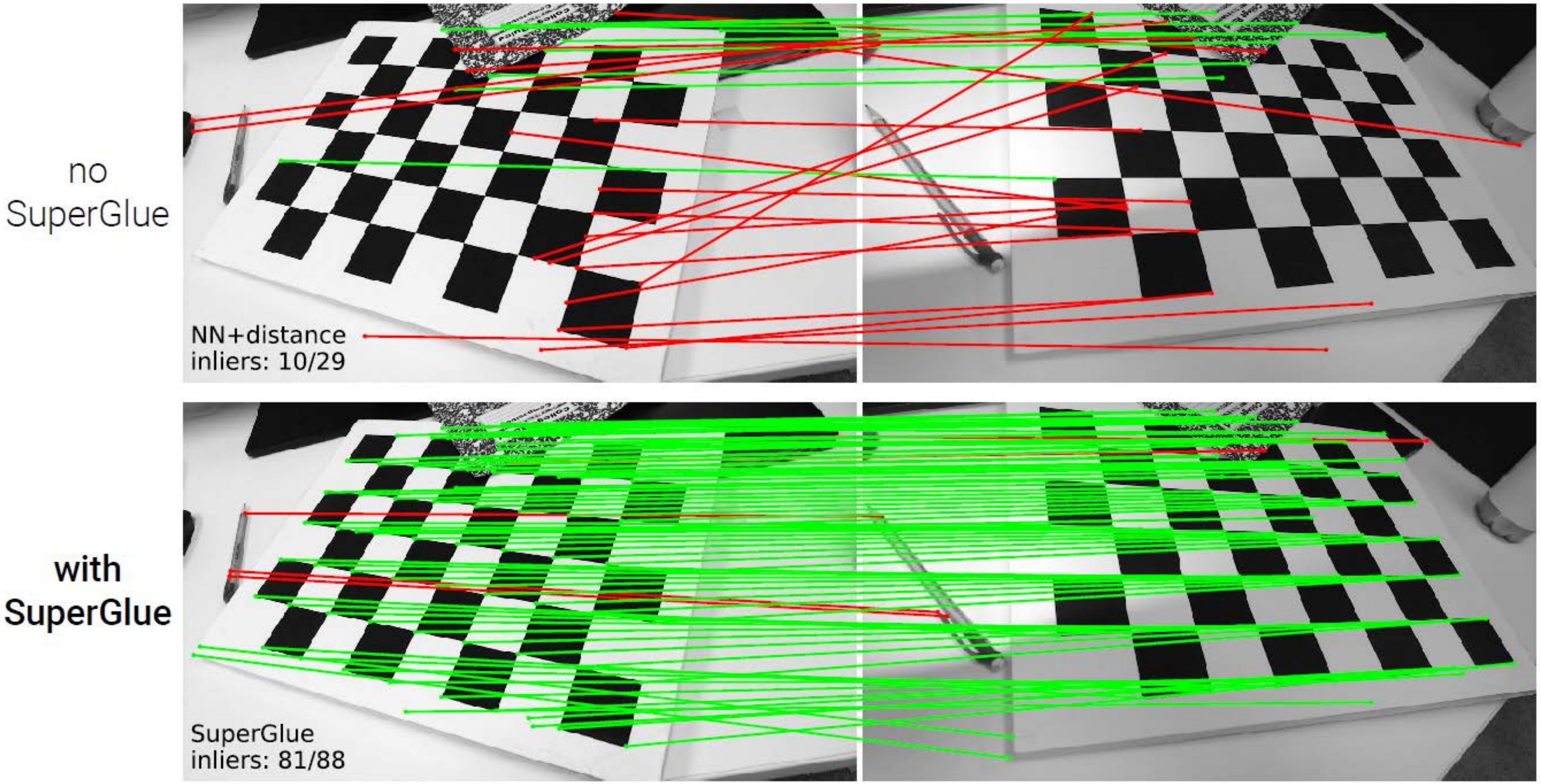
[DeTone et al, 2018]



[Yi et al, 2018]

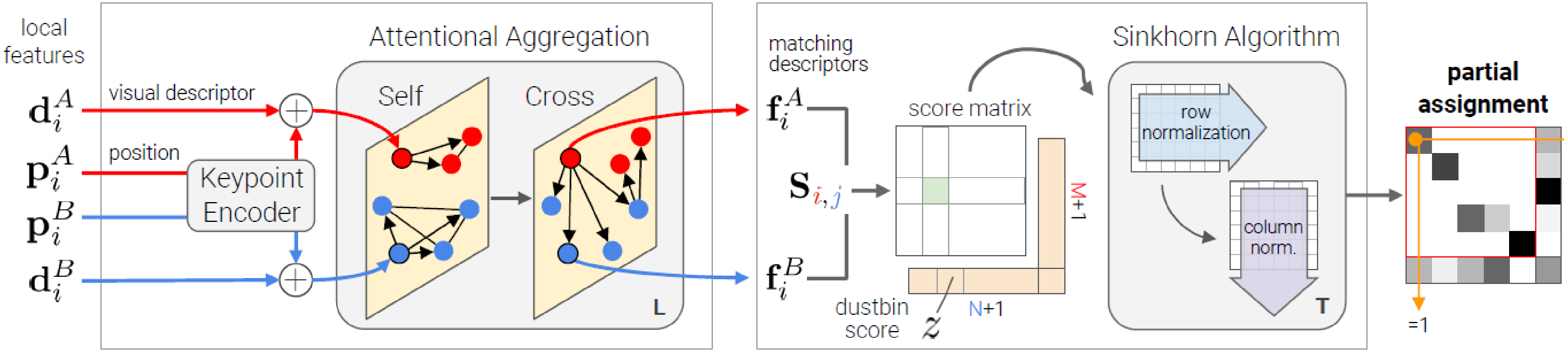
SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

The importance of context



SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

SuperGlue



A Graph Neural Network with attention

Solving a partial assignment problem

Encodes **contextual cues** & priors

Reasons about the 3D scene

Differentiable **solver**

Enforces the assignment constraints
= **domain knowledge**

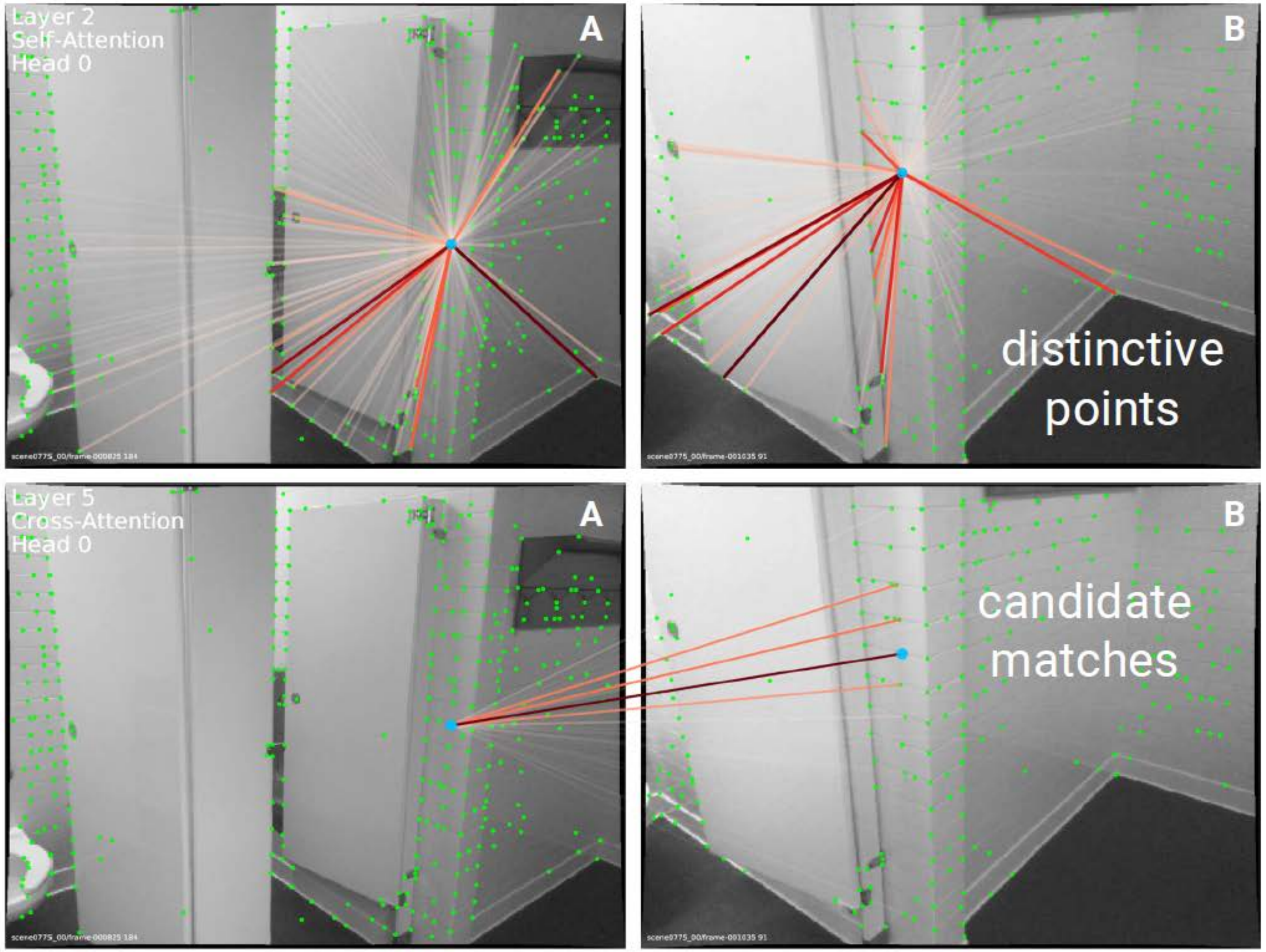
SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020. https://arxiv.org/abs/1911.11763

Self-attention and Cross-attention

Self-attention
= intra-image
information
flow

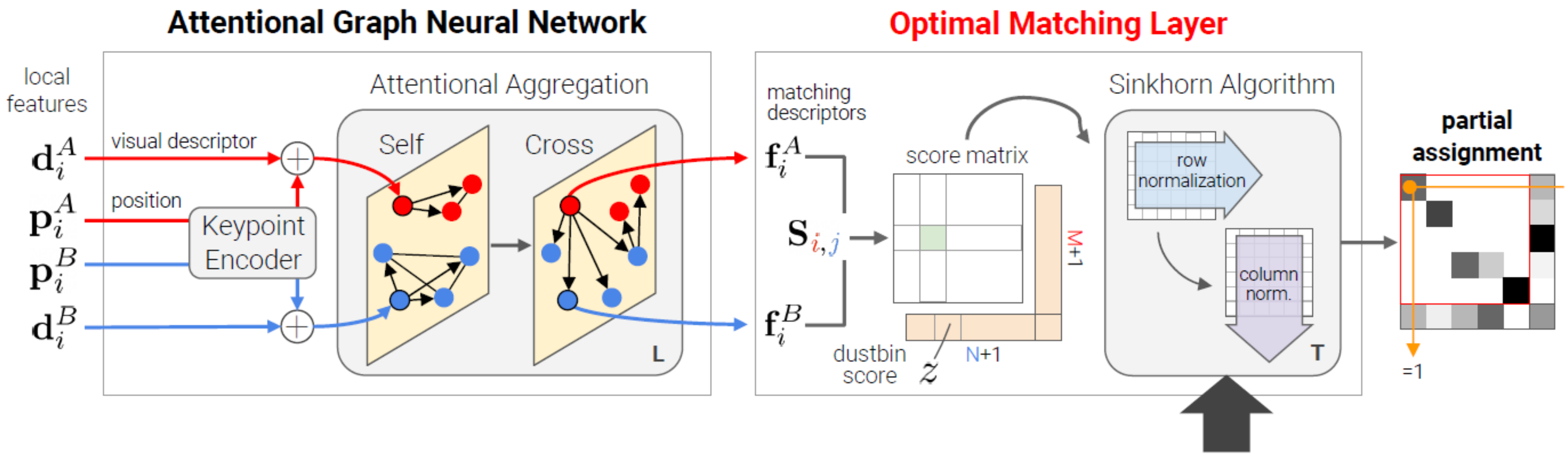
Cross-attention
= inter-image

Attention builds a
soft, **dynamic**,
sparse graph



SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

SuperGlue – Optimal Matching Layer

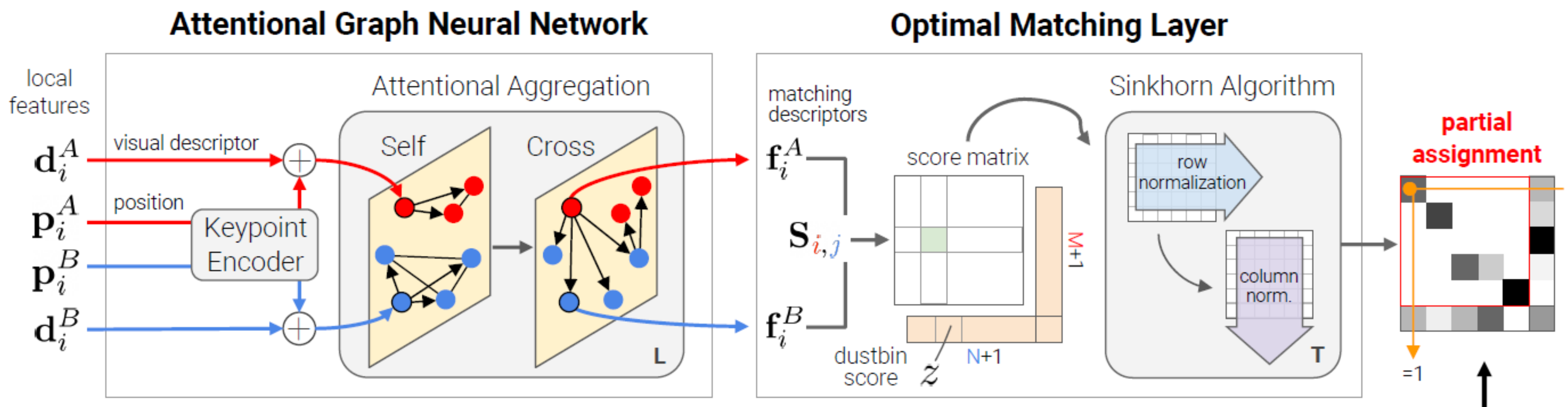


- Compute the assignment $\bar{\mathbf{P}}$ that maximizes $\sum_{i,j} \bar{\mathbf{S}}_{i,j} \bar{\mathbf{P}}_{i,j}$
- Solve an **optimal transport** problem
- With the **Sinkhorn algorithm**: differentiable & soft Hungarian algorithm

[Sinkhorn & Knopp, 1967]

SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

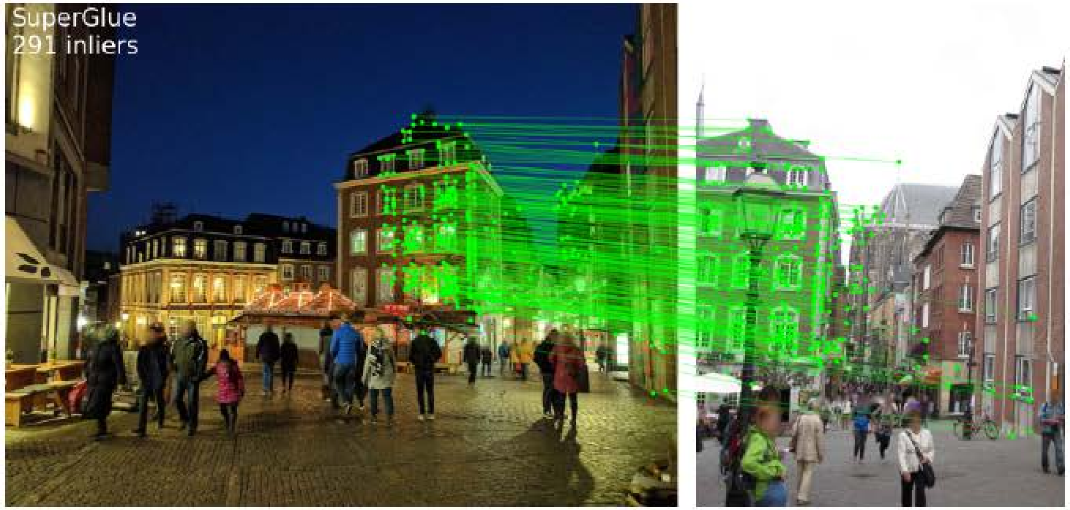
SuperGlue



- Compute **ground truth correspondences** from pose and depth
- Find which keypoints should be **unmatched**
- Loss: maximize the log-likelihood $\bar{\mathbf{P}}_{i,j}$ of the GT cells

SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

SuperGlue Example Results



SuperGlue: Learning Feature Matching with Graph Neural Networks, Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich , CVPR 2020.
https://arxiv.org/abs/1911.11763

Hierarchical Localization with hloc and SuperGlue

First place in 6 localization challenges!

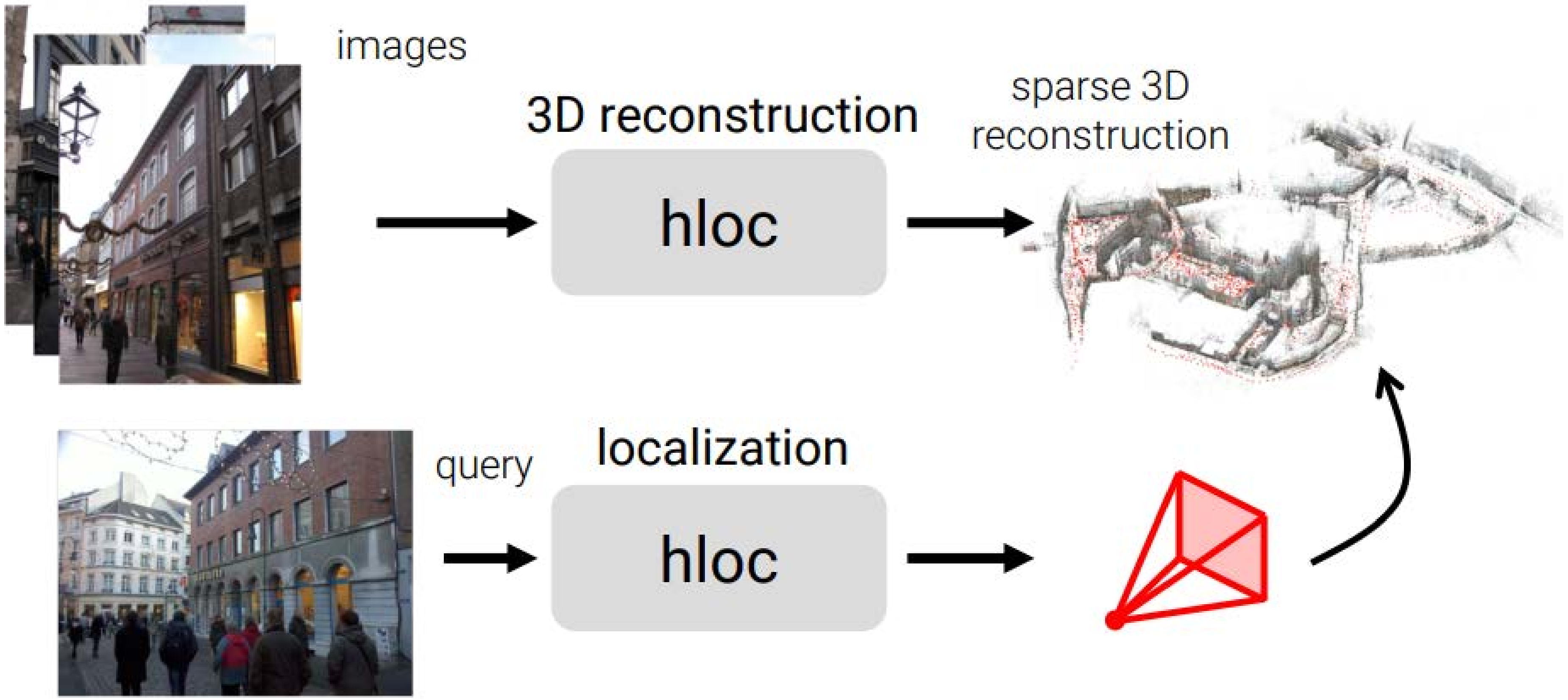
At CVPR 2020: 2 challenges, local features & handheld devices

At ECCV 2020, workshops:

- 1x Map-based Localization for Autonomous Driving
- 3x Long-Term Visual Localization under Changing Conditions

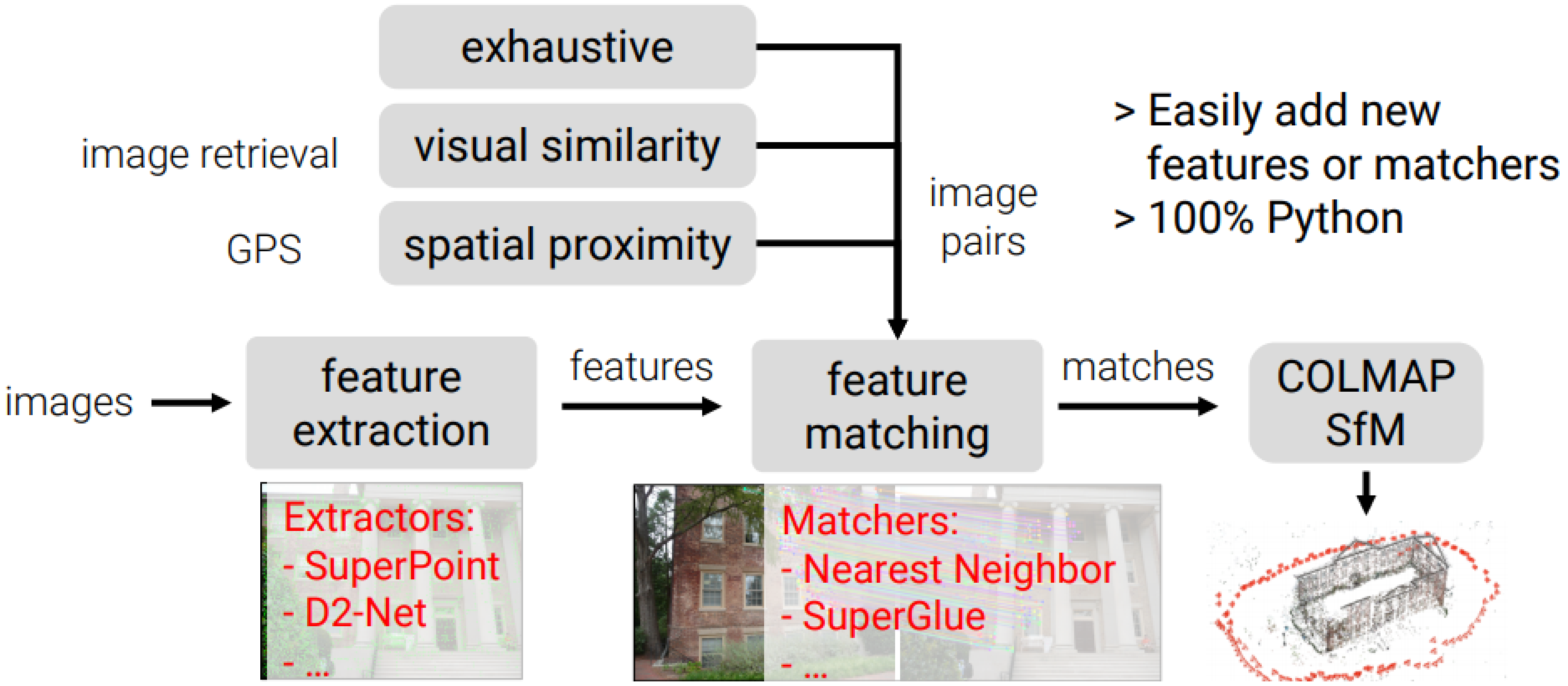
From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

Hloc – a toolbox for SFM & localization



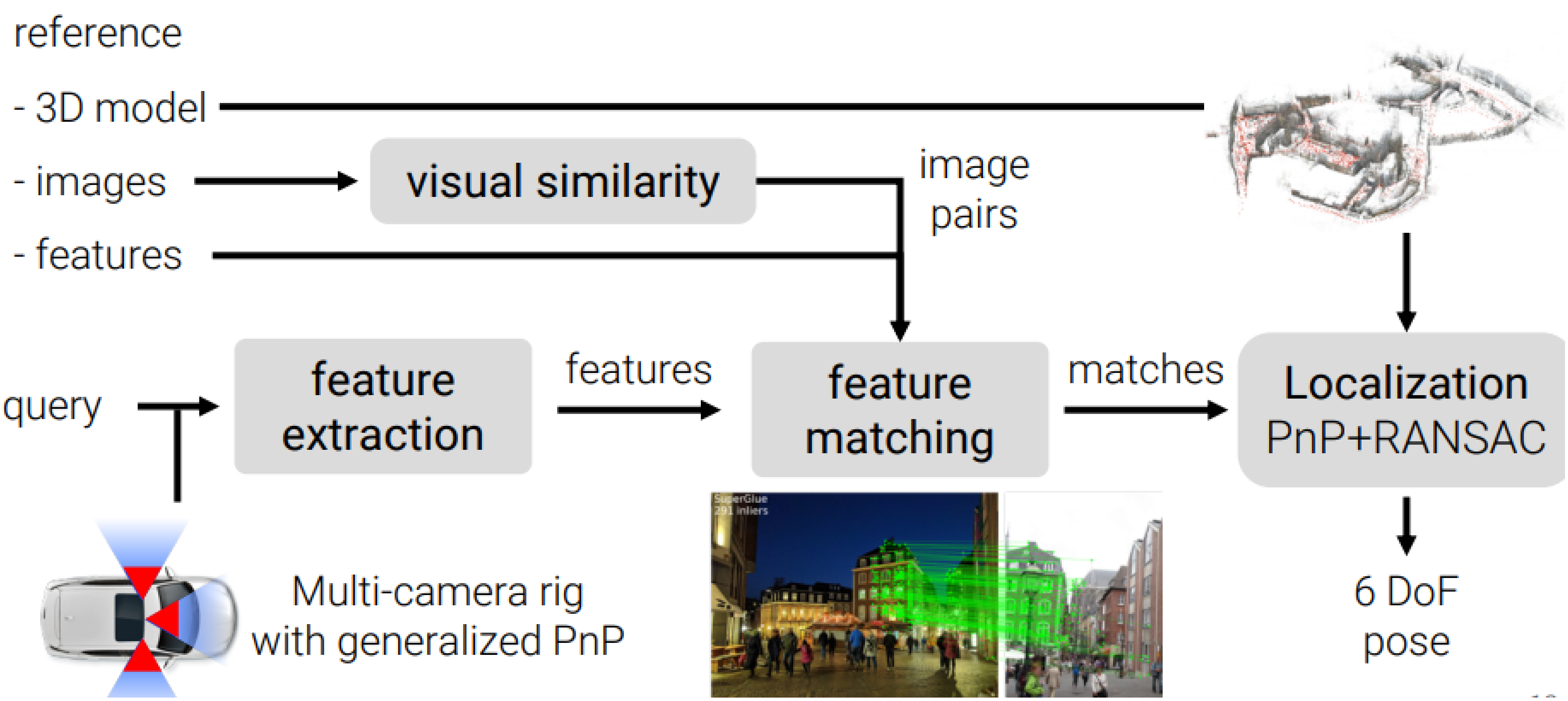
From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

Hloc - reconstruction



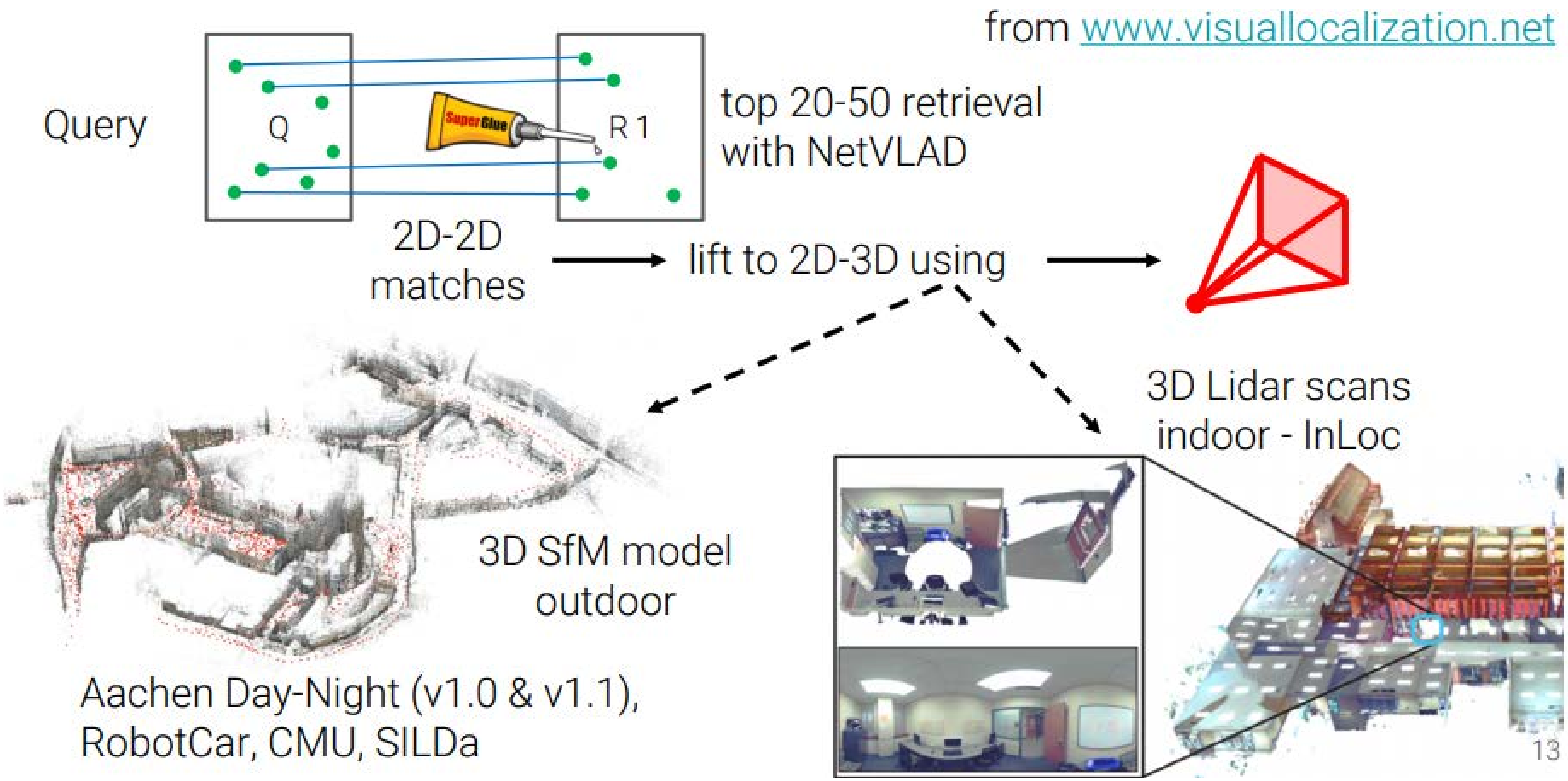
From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

Hloc- localization



From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

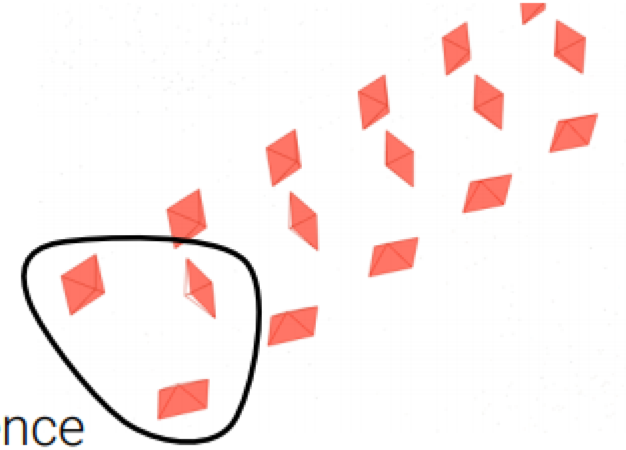
Supported datasets: Outdoor and Indoor

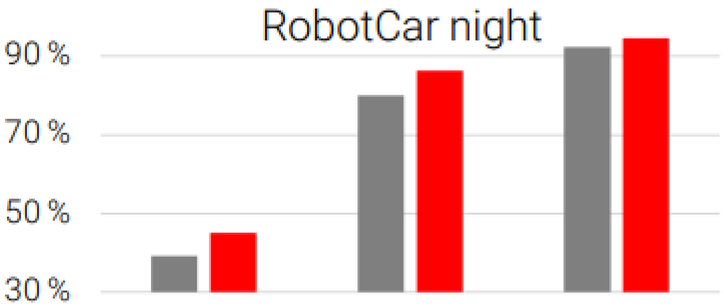


From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

Multi-camera localization for autonomous driving

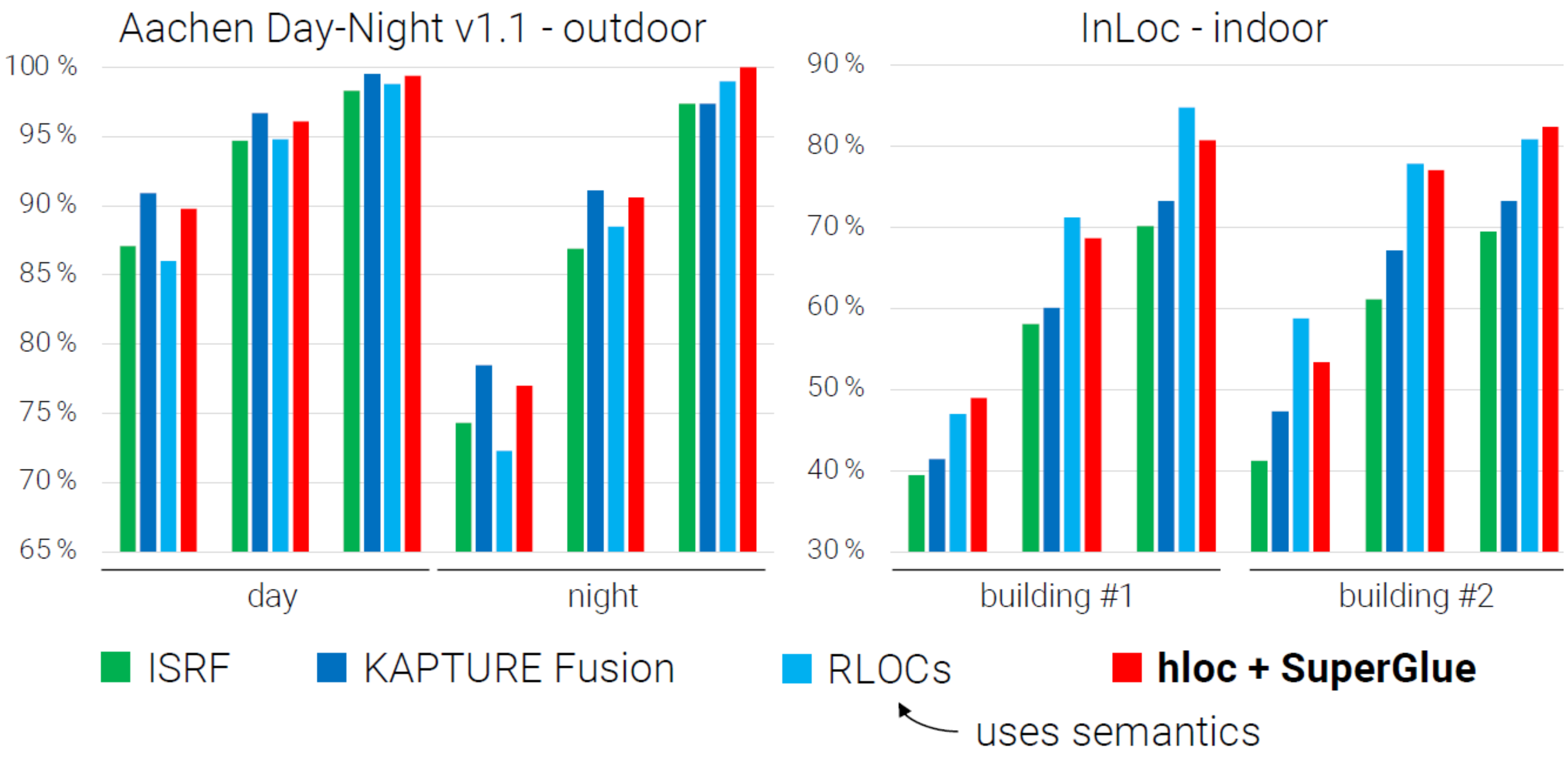
- For RobotCar, CMU and SILDa
- LO-MSAC + GP3P (RansacLib + PoseLib)
  Wald et al., ECCV 2020
  github.com/tsattler/MultiCameraPose
- Estimate rig extrinsics:
  rotation + translation averaging on reference
- Increase robustness in hard cases + better constrains the pose





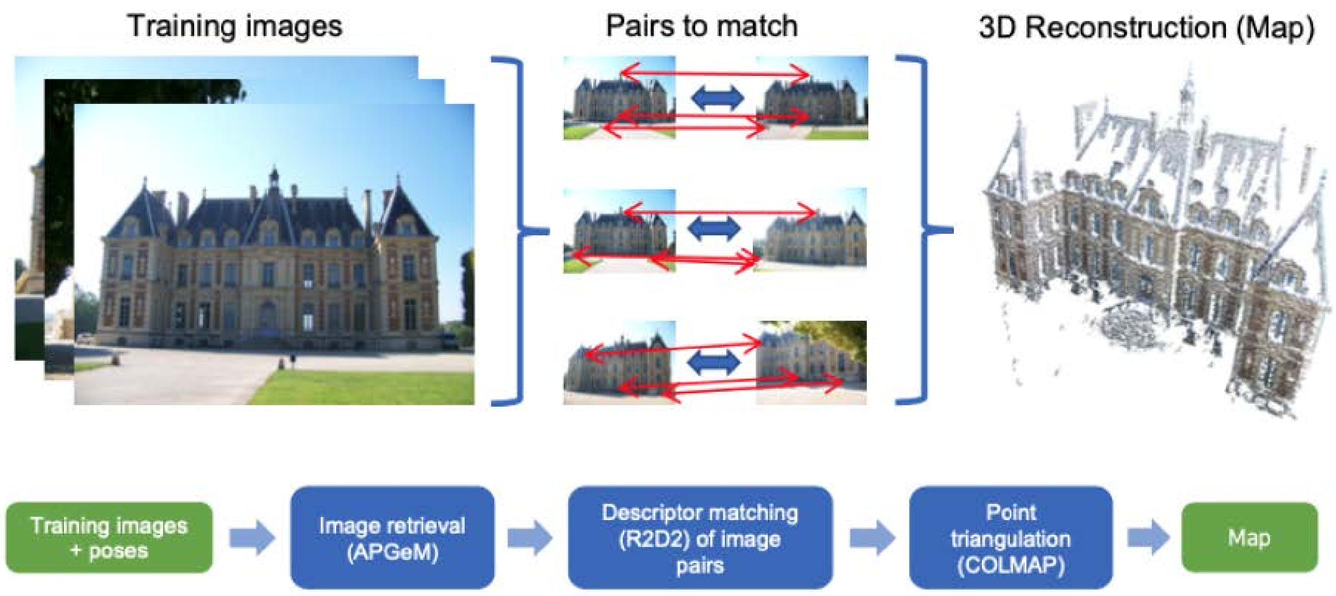
From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

ECCV 2020 Challenge Results Handheld Devices



From https://psarlin.com/assets/talks/hloc+SuperGlue_15min_ltvl_slides.pdf

Robust Image Retrieval-based Visual Localization using Kapture arXiv preprint arXiv:2007.13867



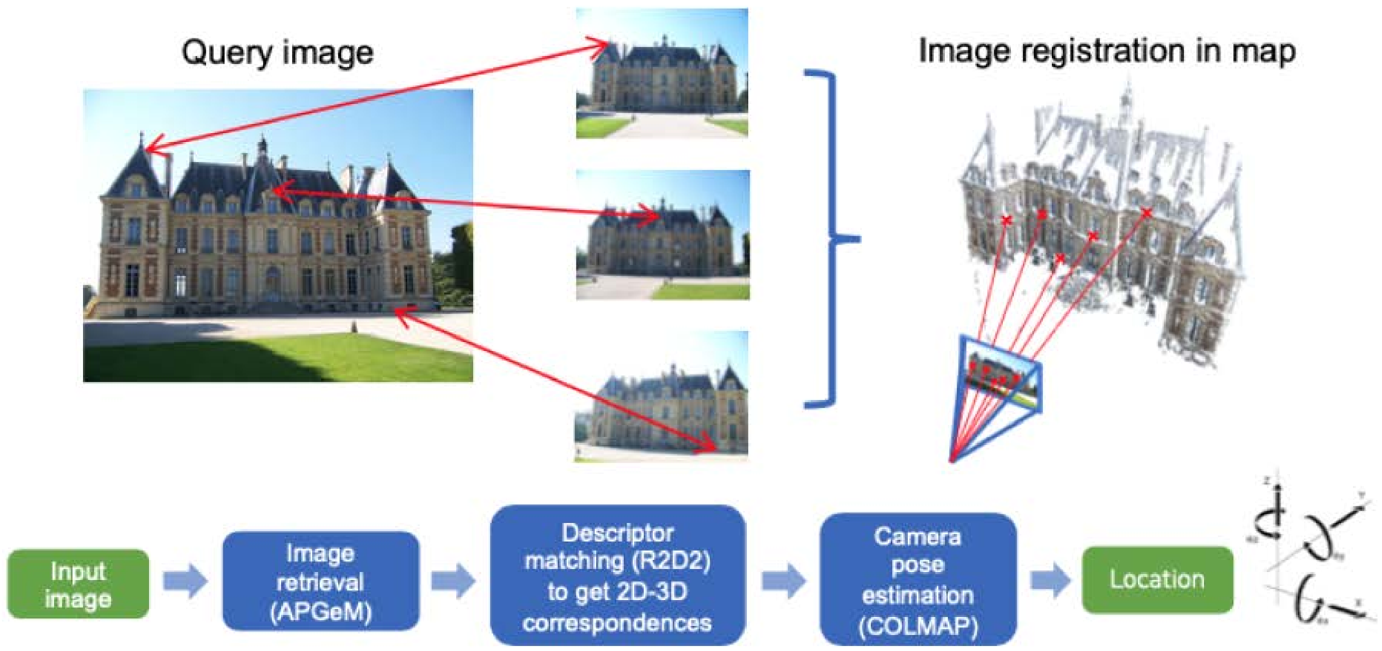
Overview of the structure from motion (SFM) reconstruction of the map from a set of training (mapping) images. Photos: Sceaux Castle image dataset

Mapping Pipeline

- Extraction of local descriptors and keypoints (e.g. R2D2) of training images
- Extraction of global features (e.g. AP-GeM) of train-ng images
- Computation of training image pairs using image retrieval
- Computation of local descriptor matches between these image pairs
- Geometric verification of the matches and point triangulation with COLMAP

- **Learning with Average Precision: Training Image Retrieval with a Listwise Loss** Jerome Revaud, Rafael S. Rezende, Cesar de Souza, Jon Almazan. ICCV 2019
- R2D2: Reliable and repeatable detector and descriptor, J Revaud, C De Souza, M Humenberger, P Weinzaepfel, Advances in Neural Information Processing Systems 32, 12405-12415

Robust Image Retrieval-based Visual Localization using Kapture: arXiv preprint arXiv:2007.13867



Overview of the localization pipeline which registers query images in the SFM map. Photos: Sceaux Castle image dataset

Localization Pipeline

- Extraction of local and global features of query images
- Retrieval of similar images from the training images
- Local descriptor matching
- Geometric verification of the matches and camera pose estimation with COLMAP

- **Learning with Average Precision: Training Image Retrieval with a Listwise Loss,** Jerome Revaud, Rafael S. Rezende, Cesar de Souza, Jon Almazan. ICCV 2019
- R2D2: Reliable and repeatable detector and descriptor, J Revaud, C De Souza, M Humenberger, P Weinzaepfel, Advances in Neural Information Processing Systems 32, 12405-12415

Conclusions

- NET-Vlad method has been adapted for a number of different retrieval methods
- Minimizing approximate AP-Precision loss gives a good gain for ranking problems
- Learning pooling choice (Max vs Sum) by Generalized mean pooling gives much better results
- Learning end-to-end makes a big difference
  - Joint learning of representation for global features (coarse search) and local features (fine alignment).
  - Enables more compact network which can run on mobile processors.
  - Cool use of teacher networks, where student is trained by two teachers.
- Attention based methods help provide context in matching. They are useful both for coarse search using semantic information and also fine alignment