Learned Contextual Feature Reweighting for Image Geo-Localization

UNC DEPARTMENT OF COMPUTER SCIENCE

Hyo Jin Kim
UNC Chapel Hill

Enrique Dunn
Stevens Institute of Technology

Jan-Michael Frahm
UNC Chapel Hill

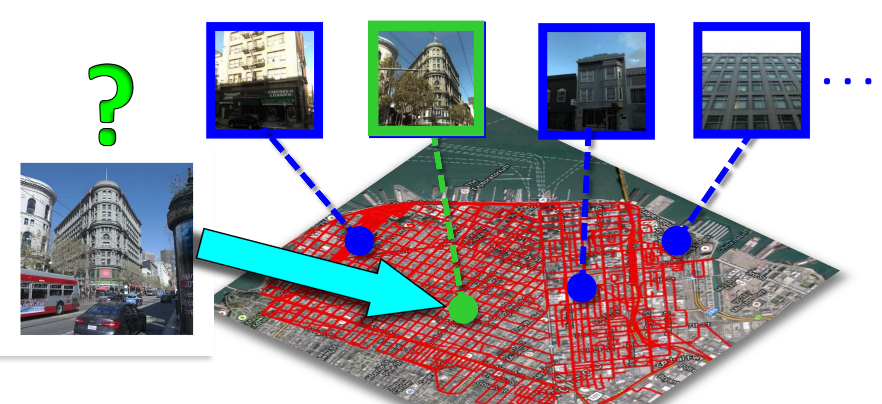
STEVENS INSTITUTE of TECHNOLOGY

http://hyojin.web.unc.edu/crn

Image Geo-Localization

Where **was this photo taken?**

➔ Find reference images (with GPS-tag) that depict the same place as the query



Challenges

- Photometric/geometric change
- **Distracting visual elements**

Applications:

- Adding/correcting GPS-tags to images
- Navigation for robots and cars
- Organizing personal photo collections

Related Work

- Feature selection and reweighting
- Focus analysis on individual local features in general

Motivation

A feature's usefulness depends largely on the context in the scene

Goal: Reweight features that are useful for image geo-localization based on the image context

Q. How to find relevant contexts?

- ➢ Defining supervised priors is limited and cumbersome
- ➢ Take advantage of **end-to-end learning**
- ➢ Network learns relevant context and weighting as it tries to minimize the geo-localization error (using only GPS-tags)

Contributions

- Propose a novel end-to-end network for learning image representations that integrates context aware feature reweighting which significantly boosts performance of the state-of-the-art.
- The proposed Contextual Reweighting Network (CRN) is fully convolutional and can be combined with any representations
- Our training pipeline only requires image geo-tags as weak supervision
- Discovery of task relevant contexts as a byproduct of training, which captures rich high level information

Contextual Reweighting Network (CRN)

Design the network architecture to capture context and weigh the contribution of each feature accordingly

$f(\text{image}) = [f_1, f_2, \dots, f_K]$

$$f_k = \sum_{l \in R} m_l \cdot a_l^k (d_l - c_k)$$

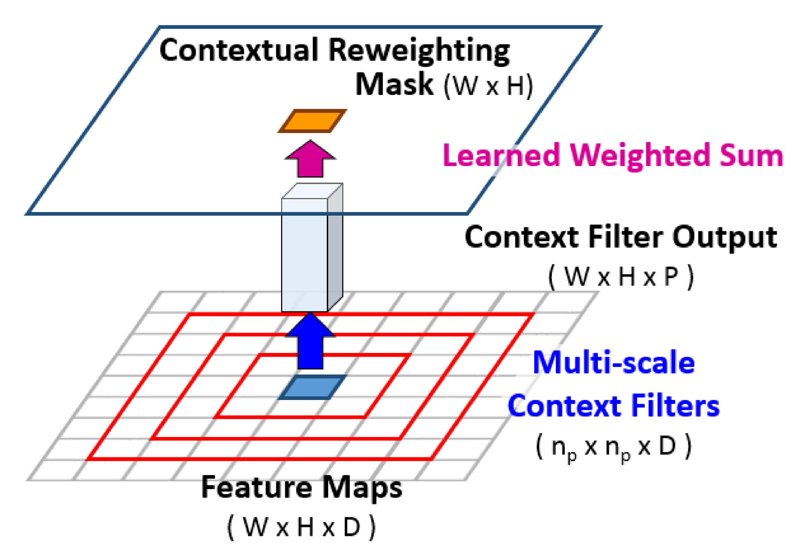
feature at spatial location l

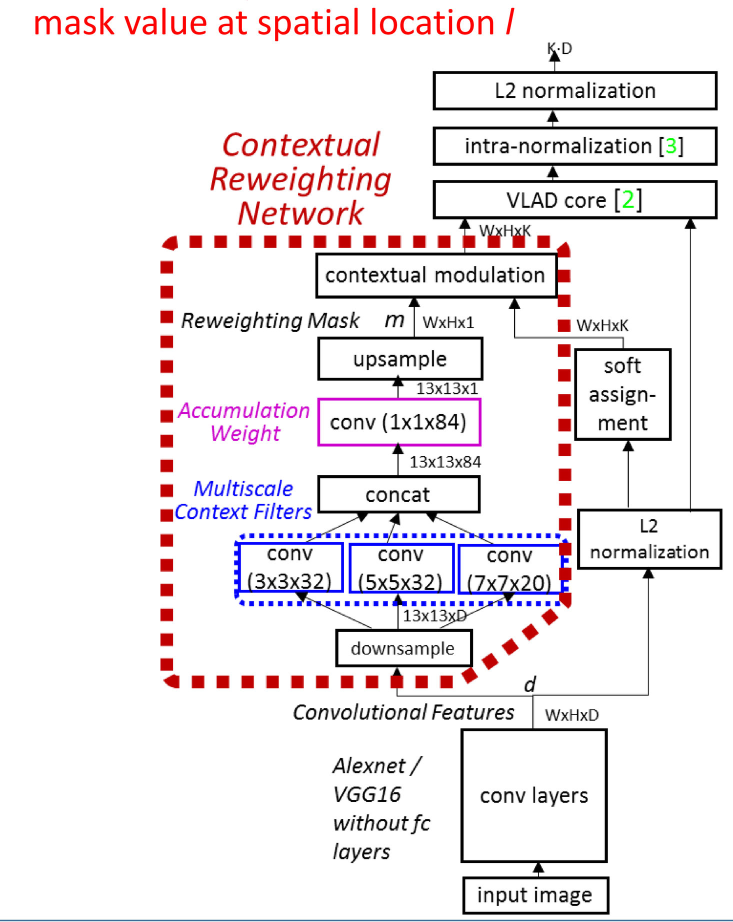
mask value at spatial location l

c_k: centroid
a_l^k: soft assignment
R: WxH spatial locations

Contextual Reweighting Network

- Shallow auxiliary network that can be used with standard representations
- Estimate the weight for a feature by multi-scale contextual information
- Each feature is reweighted, then aggregated to produce an overall representation



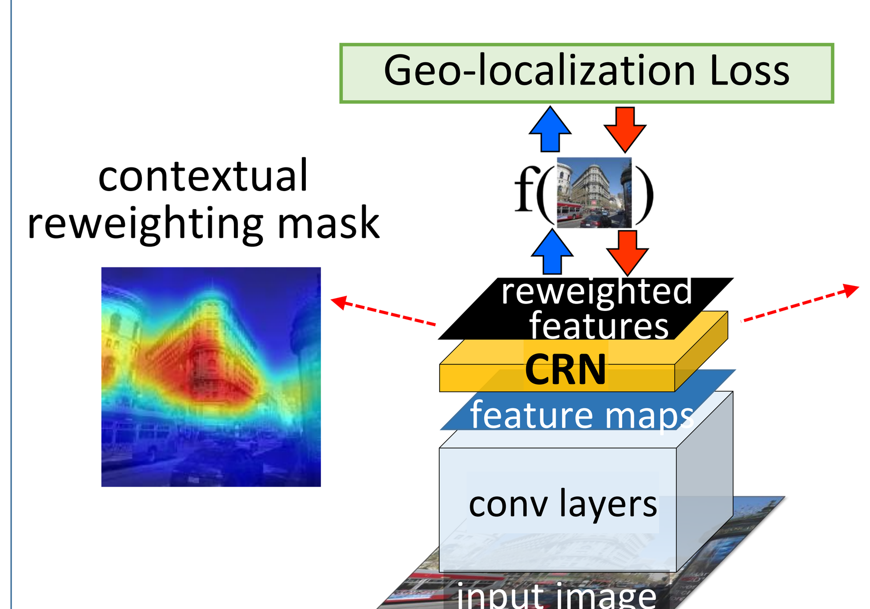


Training

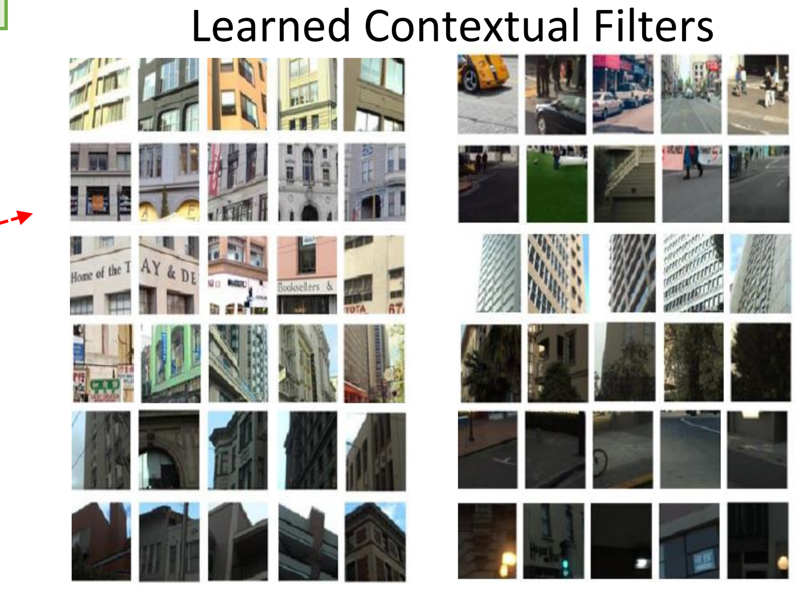
Training Objective

Triplet ranking loss for image geo-localization

$$L_f(I_t, I_r^+, I_r^-) = \max(0, \|f(I_t) - f(I_r^+)\|_2 - \|f(I_t) - f(I_r^-)\|_2)$$



Learned Contextual Filters



Automatic Training Triplet Generation

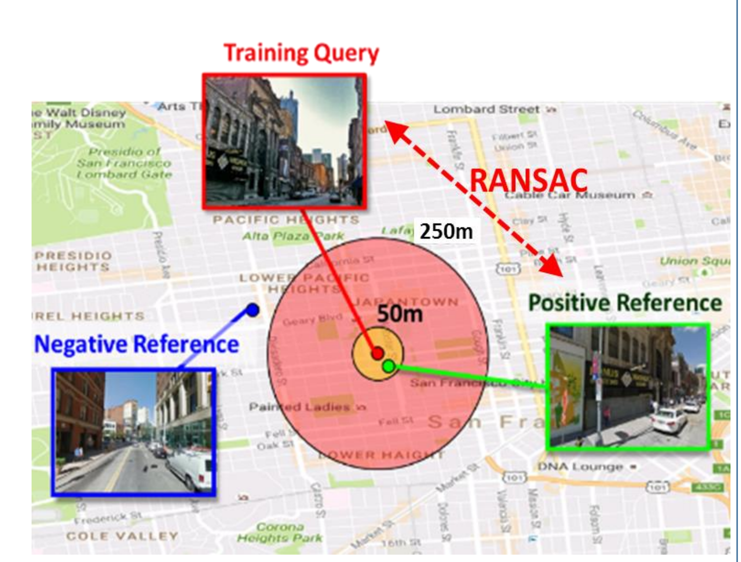
Input: A set of images with GPS-tags (e.g. Flicker images), Reference images
Output: Training triplets with training query, positive, and negative (I_t, I_r^+, I_r^-)

➢ **Positive Determination**

- Neighboring images in geographical space that pass ***geometric verification***
- Inlier-based ROI cropping for noise removal and data augmentation

➢ **Within-Batch Hard Negative Mining**

- Images within the batch that are closest to the training query in the feature space, but are far away geographically

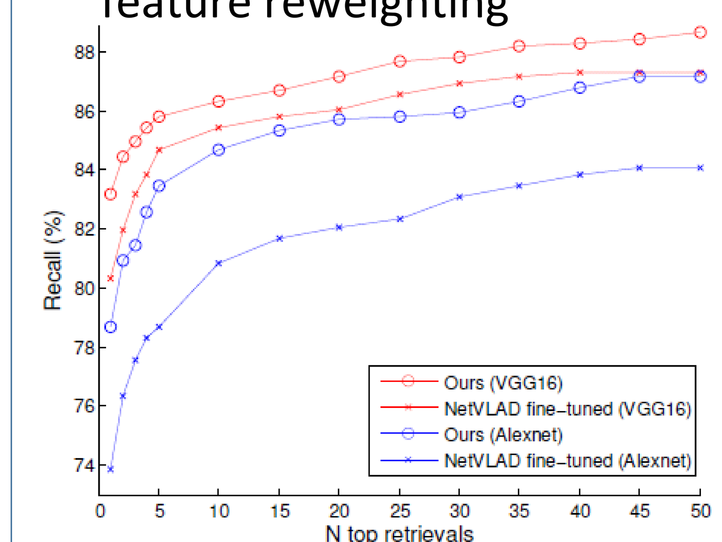


Experiments

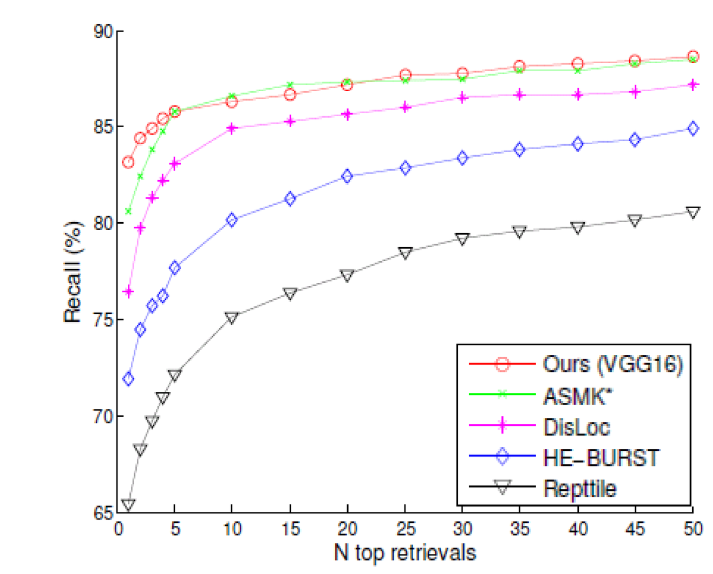
Image Geolocalization in San Francisco 1.2M Benchmark (Chen et al.)

- Query Images: 803 Hand-held mobile phone camera images
- Reference Images: 1.2M images taken from car-mounted wide angle camera
- Training query: 36K images from Flickr, Google Research data, Subset of references (144K triplets)
- Both proposed and NetVLAD trained using the same training pipeline

▪ With and without contextual feature reweighting



▪ Comparison with the state-of-the-arts



▪ Recall at Top 1 retrieval

Method	% Correct
Ours (VGG16)	**83.2**
ASMK* [54]	80.6
NetVLAD [2] fine-tuned (VGG16)	80.3
Ours (AlexNet)	**78.7**
DisLoc [4]	74.6
NetVLAD [2] fine-tuned (AlexNet)	73.9
HE-BURST [54]	71.9
Repttile [58]	65.4
NoGPS [10]	41.2
tf-idf [58]	23.2

▪ Comparison with CroW [Kalantidis et al. '16]

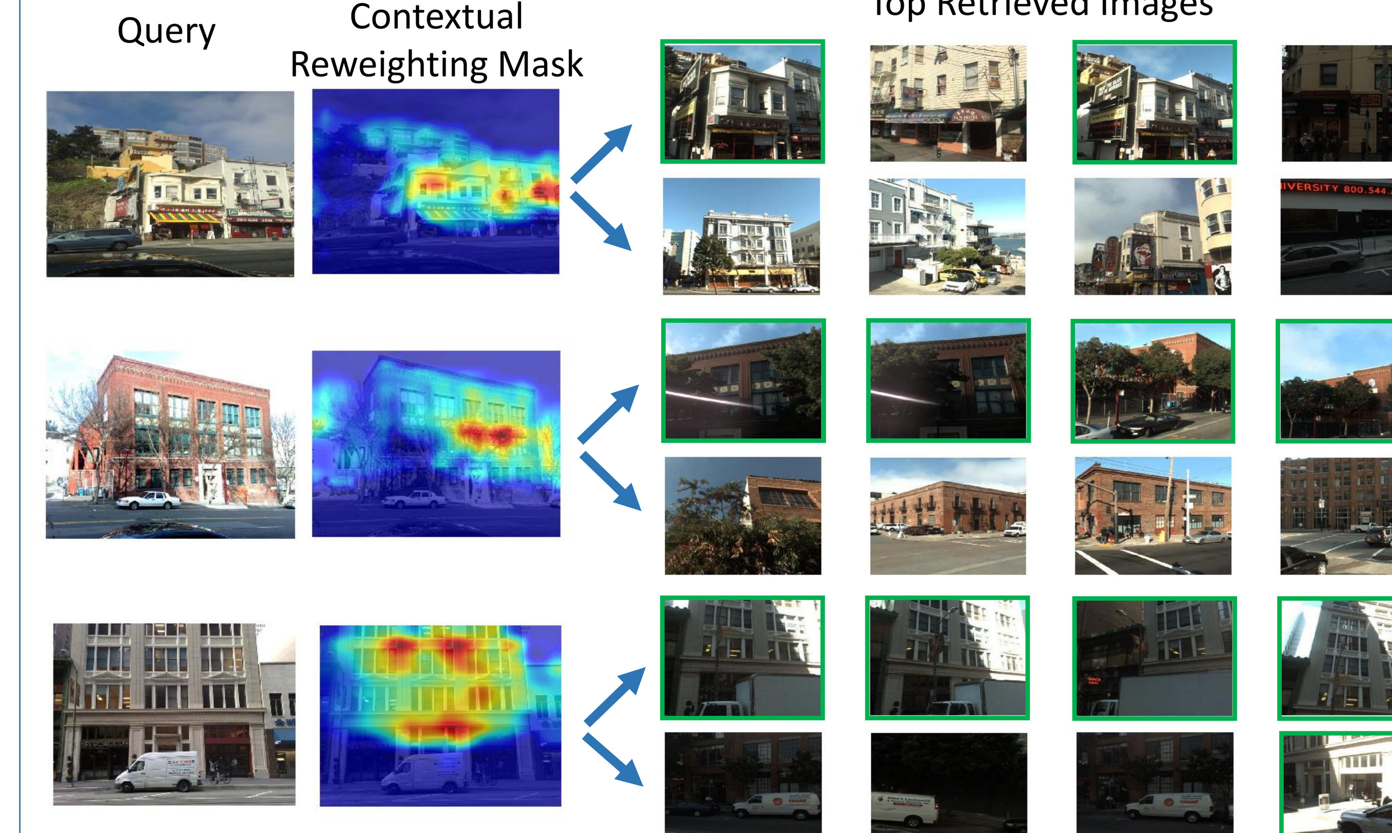
CroW creates spatial weighting as L2-norm of features s.t.features with high activation are emphasized

	top-1	top-5	top-10	top-25
CRN+NetVLAD (V)	**83.2**	**85.8**	**86.3**	**87.7**
CroW+NetVLAD (V)	80.1	84.3	85.3	86.5
CRN+NetVLAD (A)	**78.7**	**83.4**	**84.7**	**85.8**
CroW+NetVLAD (A)	74.1	79.3	80.8	82.3

Query

Contextual Reweighting Mask

Top Retrieved Images



CRN+NetVLAD

Vanilla NetVLAD

CRN+NetVLAD

Vanilla NetVLAD

CRN+NetVLAD

Vanilla NetVLAD

Tokyo 24/7 and Pittsburgh 250k

We used the same training, testing splits of NetVLAD paper
For NetVLAD, we show the recalls reported by its authors

data	set	method	top-1	top-5	top-10
Tokyo 24/7 [57]	all	**Ours**	**75.2**	**83.8**	**87.3**
		NetVLAD	71.8	82.5	86.4
	sunset /night	**Ours**	**66.7**	**76.7**	**81.9**
		NetVLAD	61.4	75.7	81.0
Pittsburgh 250K [58]	test [2]	**Ours**	85.5	**93.5**	**95.5**
		NetVLAD	**86.0**	93.2	95.1

Oxford Buildings 5k and 105k

No training, crop of ROI, or spatial reranking

	Oxford 5K [40]			Oxford 105K [40]	
Method	Ours	NetVLAD [2]		Ours	NetVLAD[2]
Dim \ Train	SF	SF	PGH	SF	SF
16384	**0.704**	0.683	-	**0.685**	0.664
8192	**0.699**	0.682	-	**0.680**	0.660
4096	**0.692**	0.672	0.691	**0.671**	0.651
2048	**0.683**	0.660	0.677	**0.662**	0.633
1024	0.667	0.650	**0.669**	**0.644**	0.625
512	0.645	0.626	**0.656**	**0.622**	0.598
256	**0.642**	0.608	0.625	**0.617**	0.579
128	**0.615**	0.569	0.604	**0.586**	0.540

Acknowledgement Supported by the Intelligence Advance Research Projects Activity (IARPA) via Air Force Research Laboratory (AFRL), contract FA8650-12-C-7214. The U.S Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government. The authors would also like to thank Relja Arandjelovi´c and Akihiko Torii for providing data, code, and sharing insights, and Alex Berg for helpful discussions.