



Large-scale urban visual geo-localization.

C. Masone, G. Berton



Dr. Carlo Masone
Politecnico di Torino

Urban visual geo-localization

principles and tips

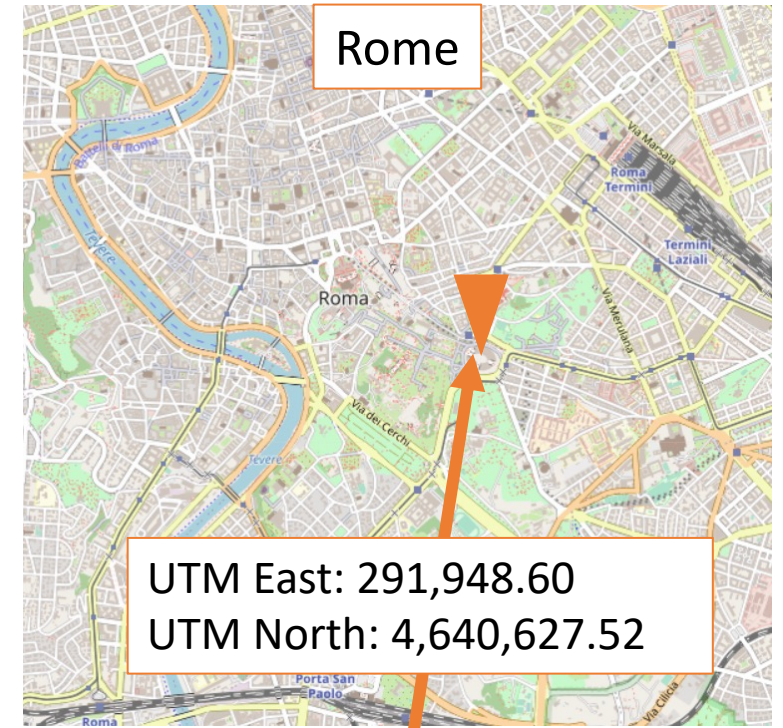
Urban visual geo-localization

Introduction and general
concepts

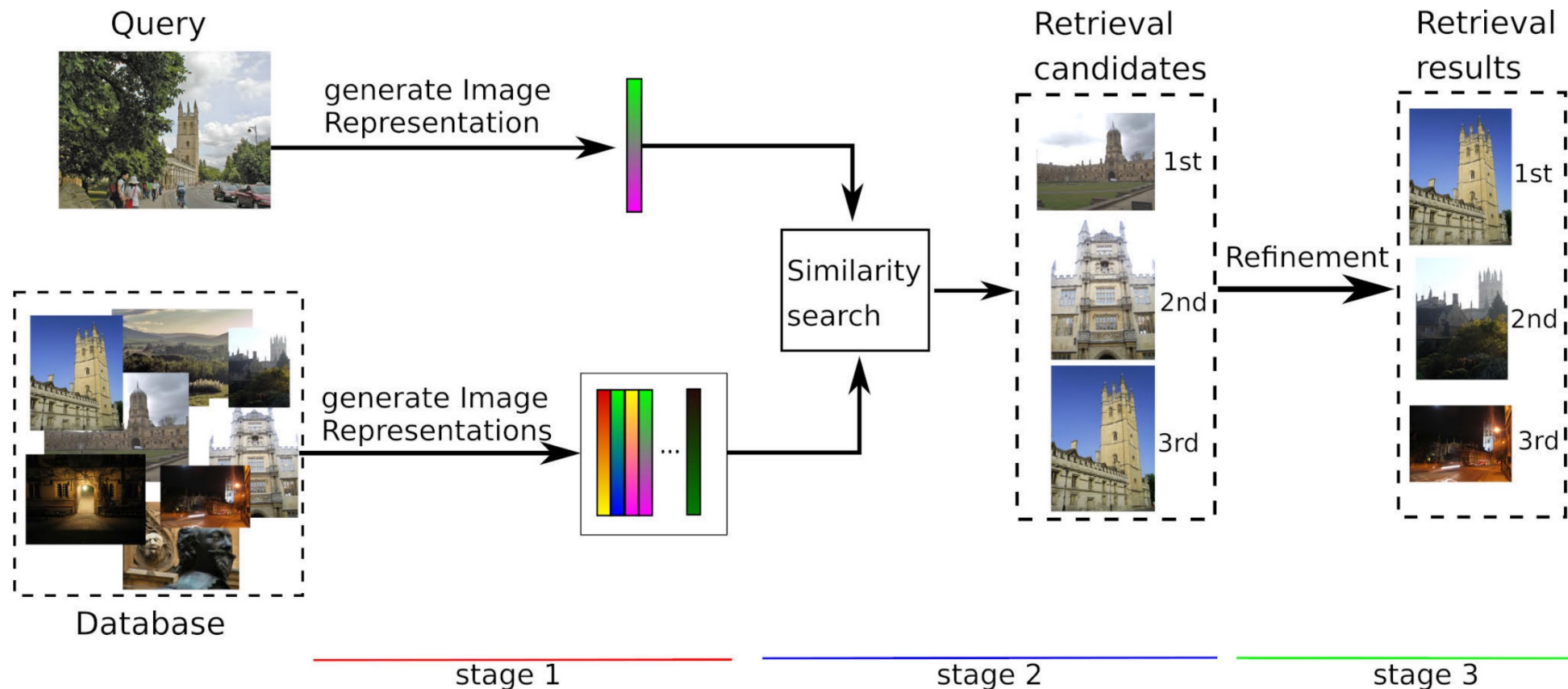


Urban visual geo-localization

- In visual geo-localization (visual place recognition) we want to estimate the coordinates (e.g., GPS or UTM) where an image was taken.
- In urban applications, the expected accuracy is usually in the range of meters (typically 25 meters)



Visual geo-localization as a retrieval problem




	Vanilla	Resize (80%)	Data augm. (brightness = 2)	Pred. refinement (<i>nearest crop</i>)	PCA (2048)	CRN [37]
R@1	63.4	64.3	68.6	67.0	56.6	68.8

Table 1. Example of how results can be influenced by little train or test time changes to the VG pipeline. Recall@1 for a ResNet-18 with NetVLAD trained on Pitts30k and tested on Tokyo24/7.

An example

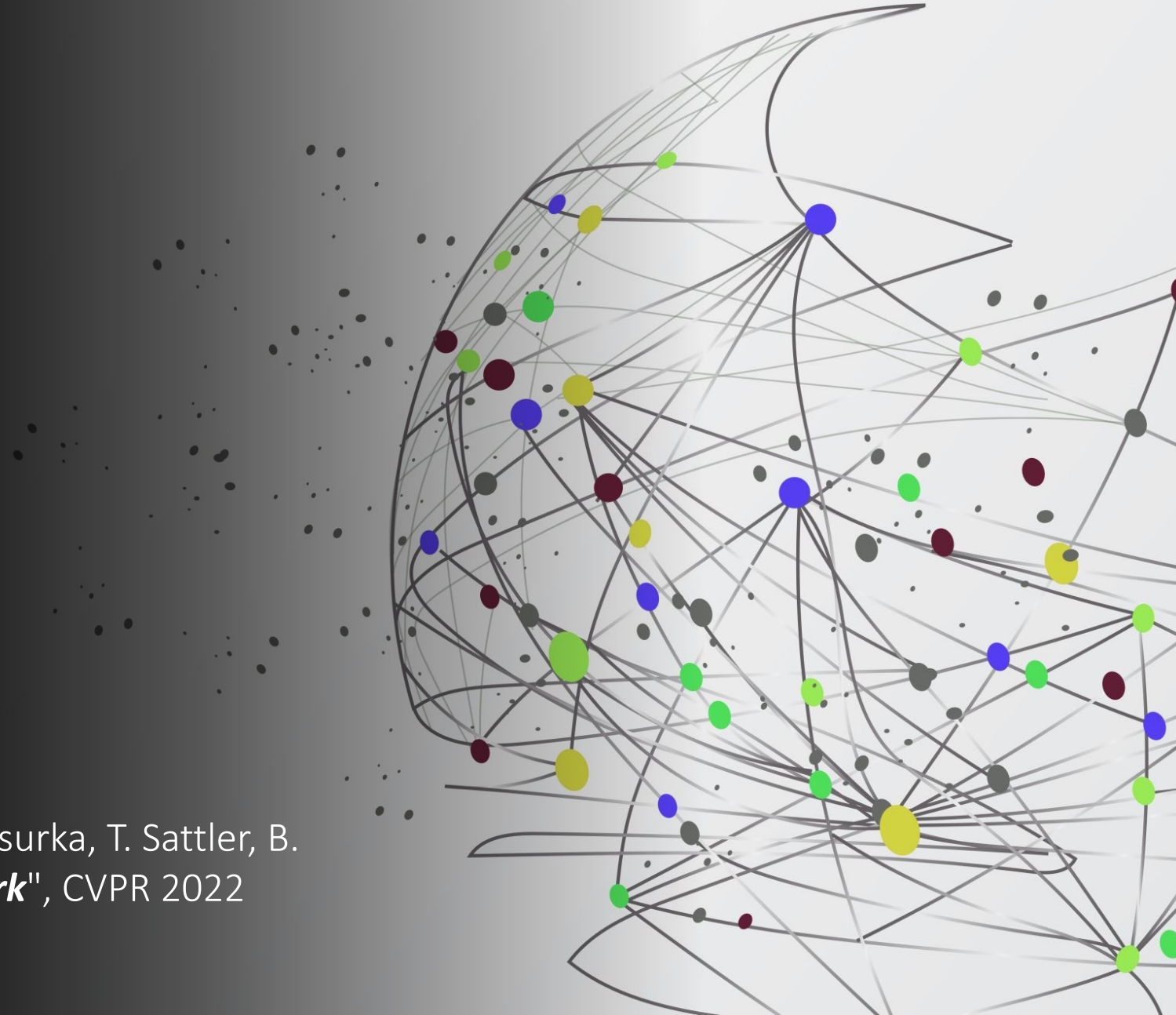
In a visual geo-localization pipeline, there are many factors and engineering choices that can significantly affect the results



Deep image visual geo- localization benchmark

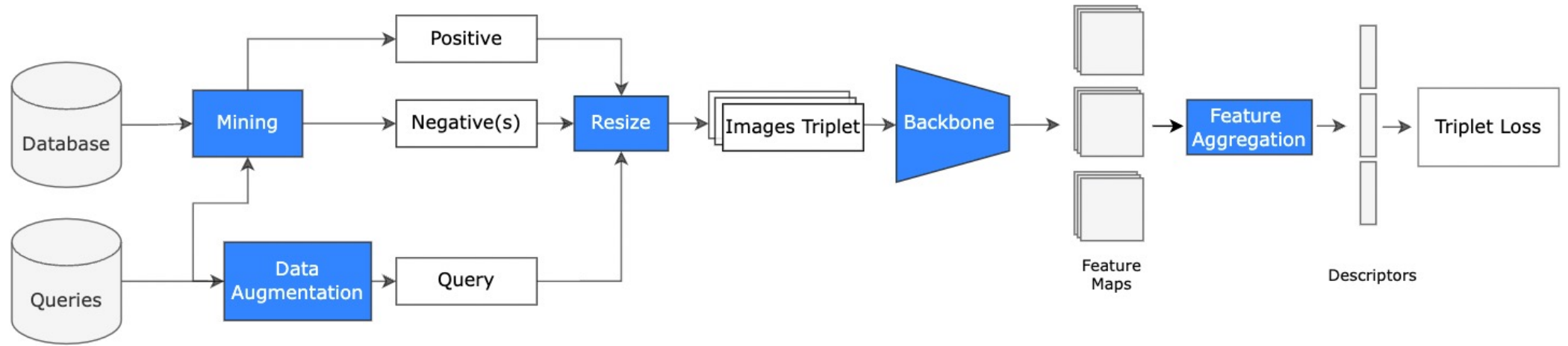
Tips and practical lessons

G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, B. Caputo, "***Deep visual geo-localization benchmark***", CVPR 2022



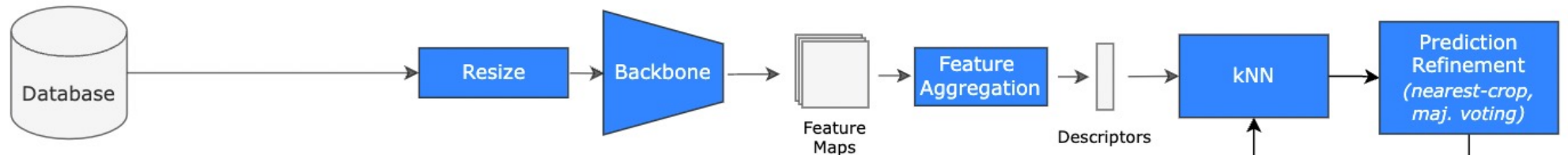
A visual geo-localization pipeline

Train Time Scenario

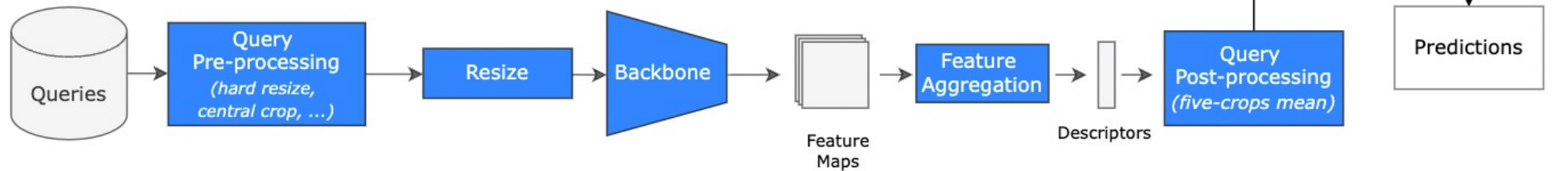


Test Time Scenario

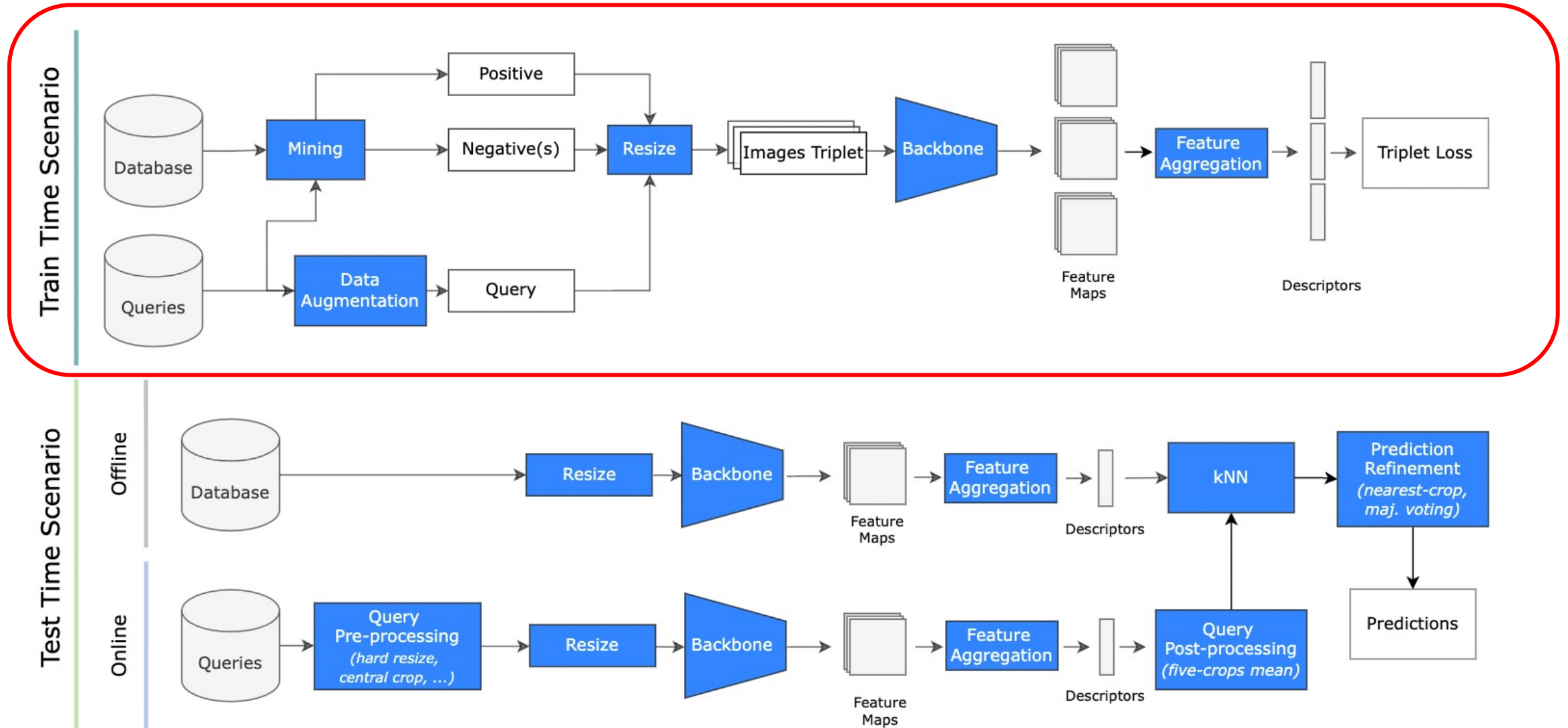
Offline



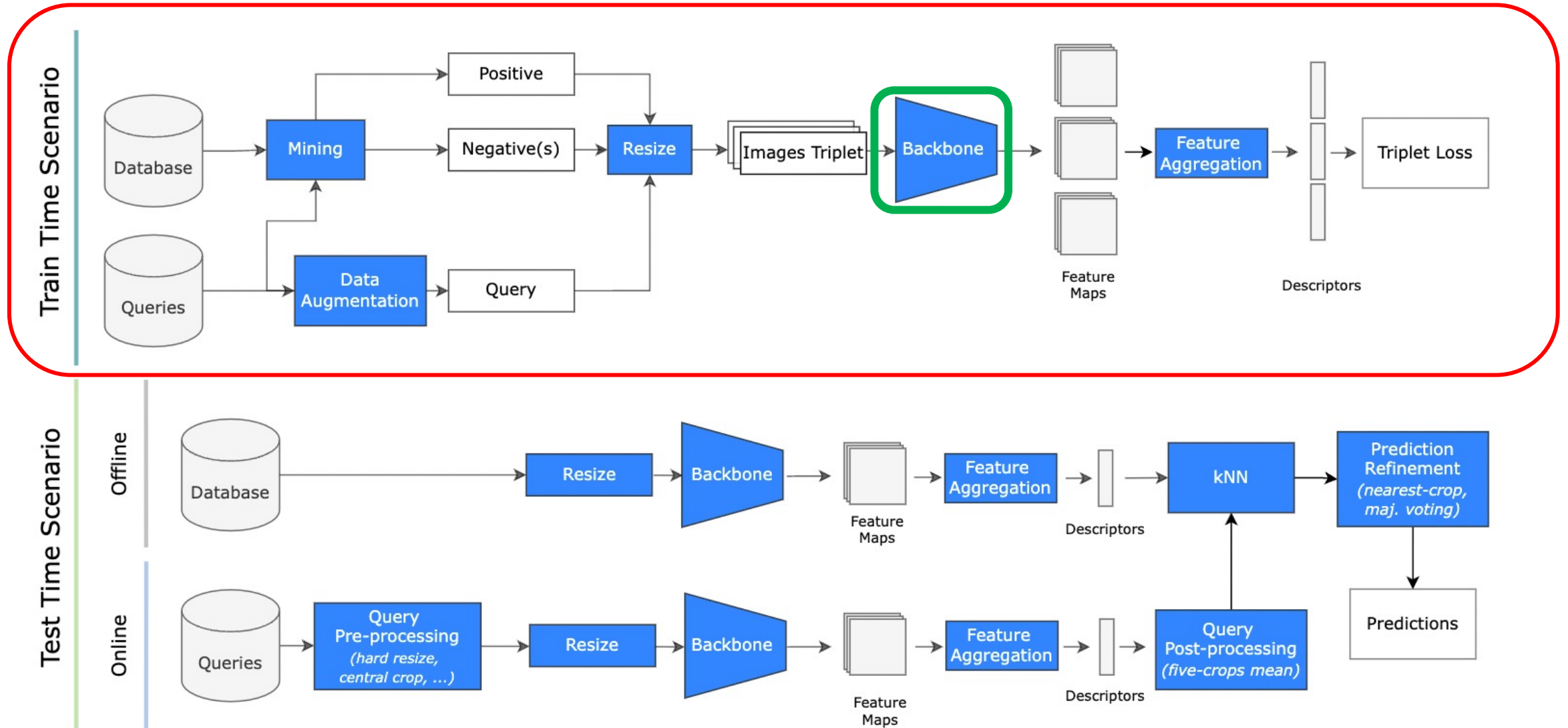
Online



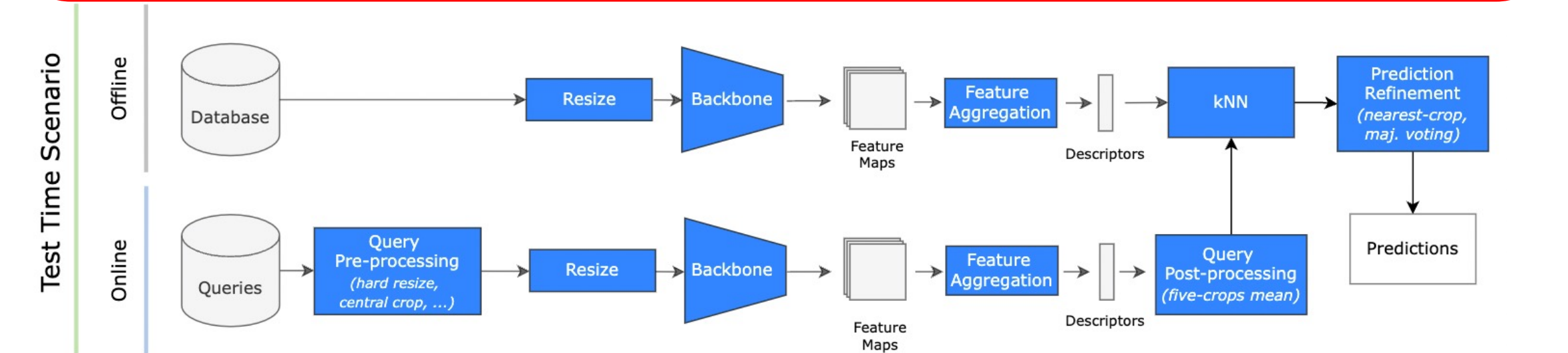
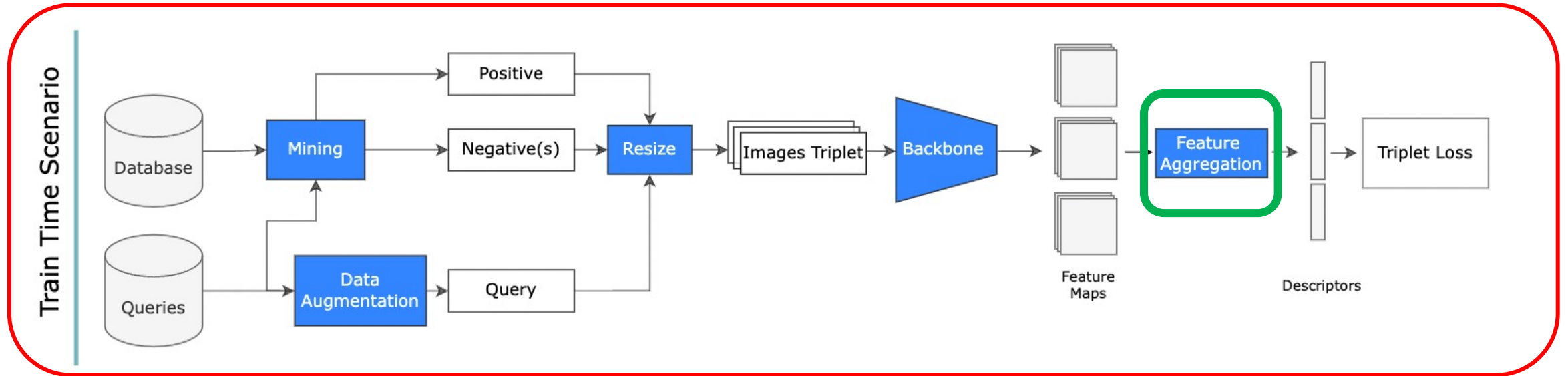
A visual geo-localization pipeline



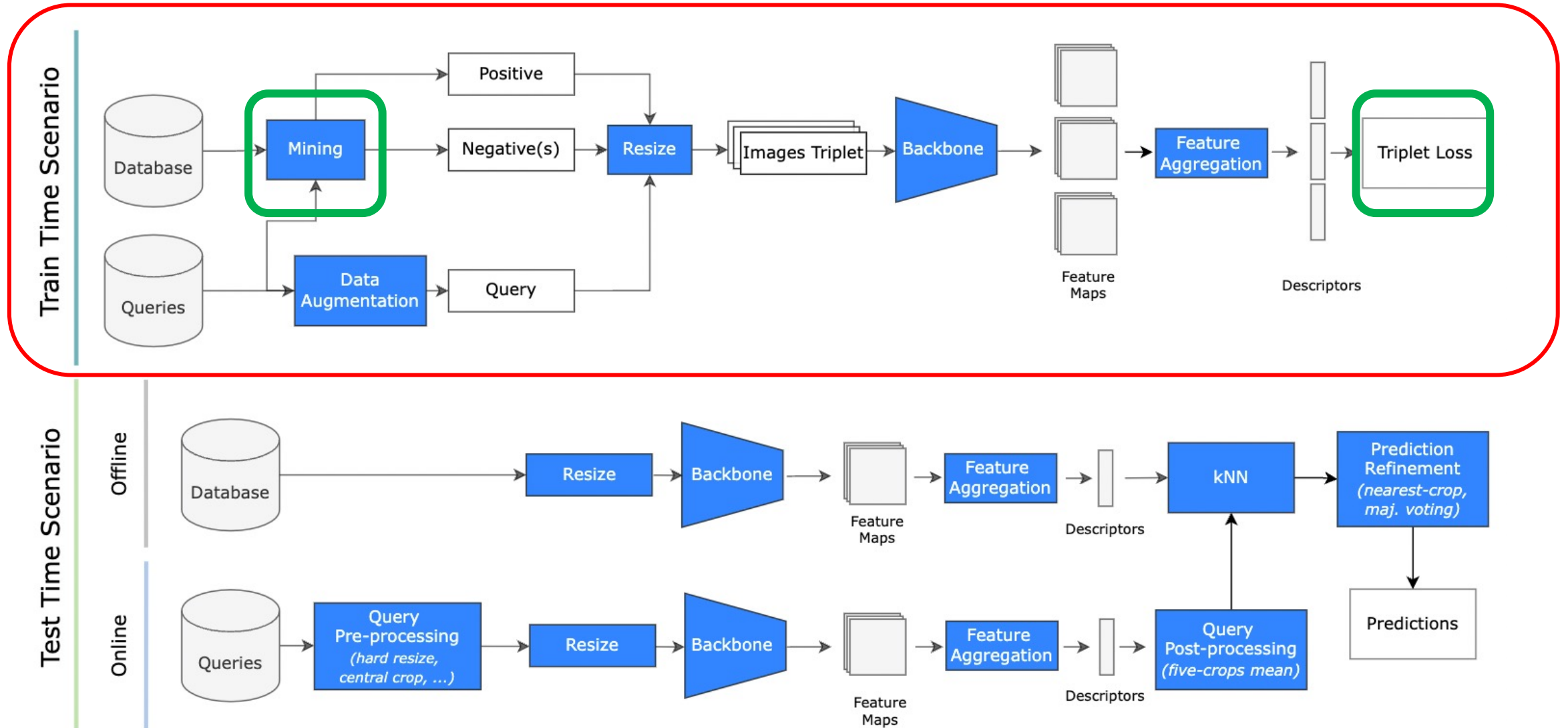
A visual geo-localization pipeline



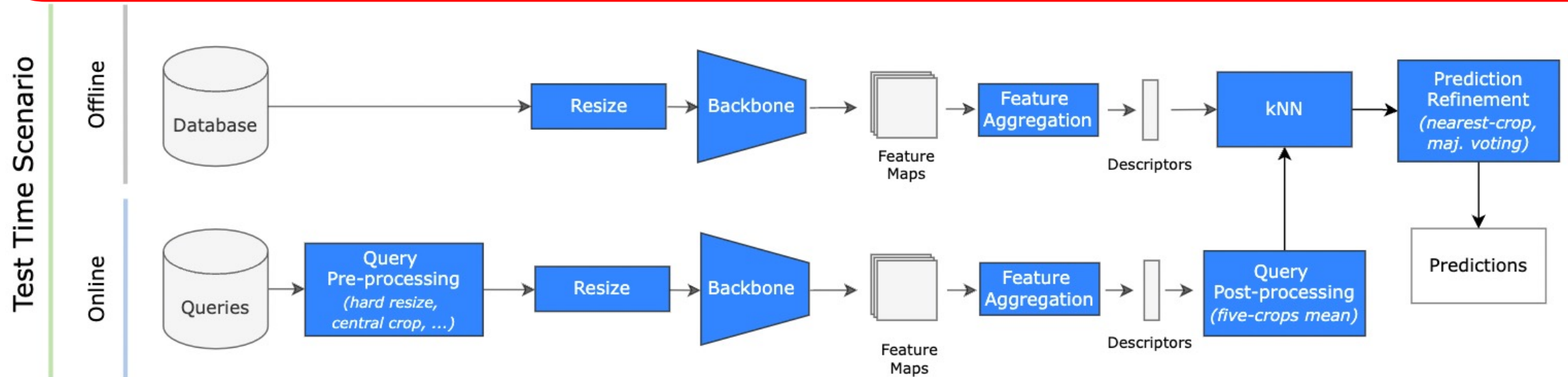
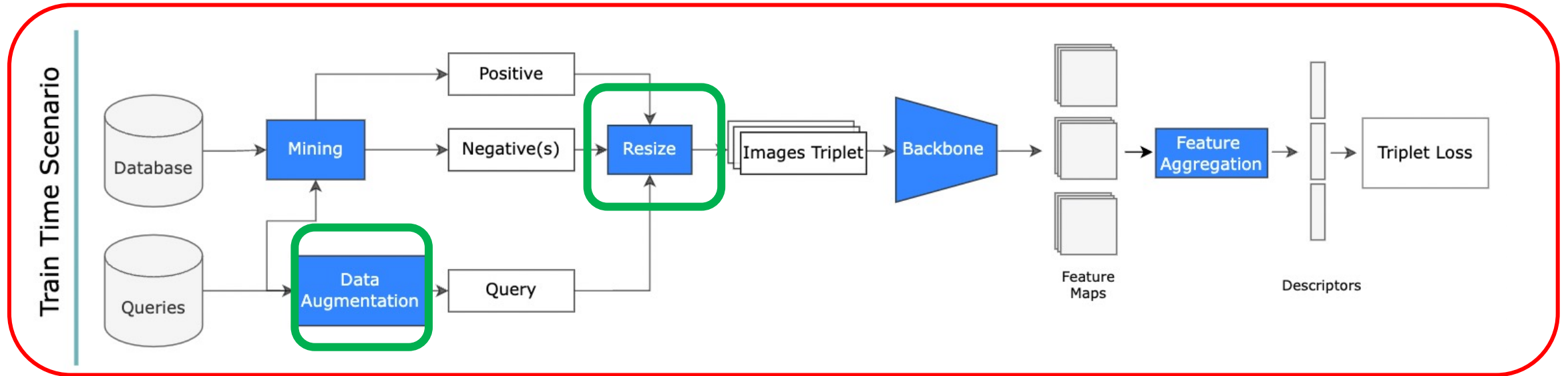
A visual geo-localization pipeline



A visual geo-localization pipeline

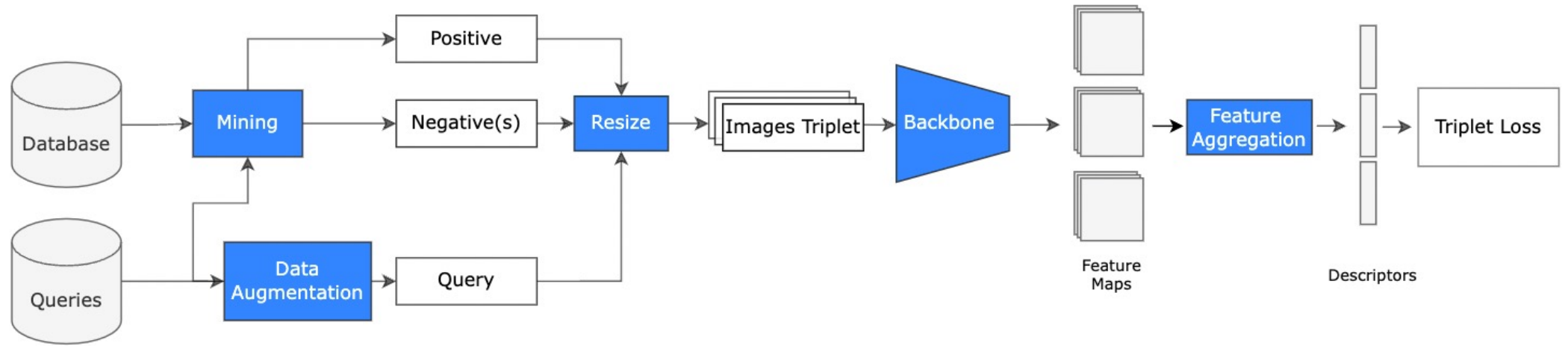


A visual geo-localization pipeline



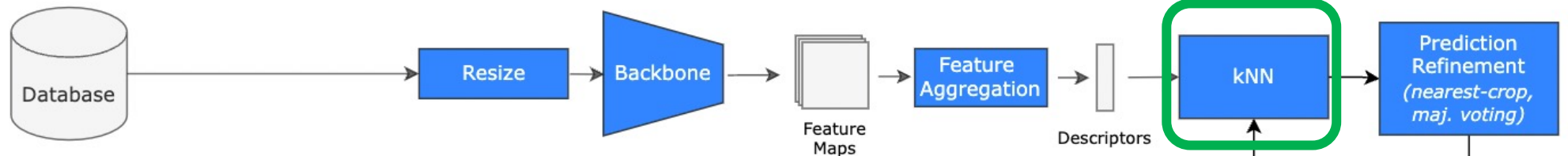
A visual geo-localization pipeline

Train Time Scenario

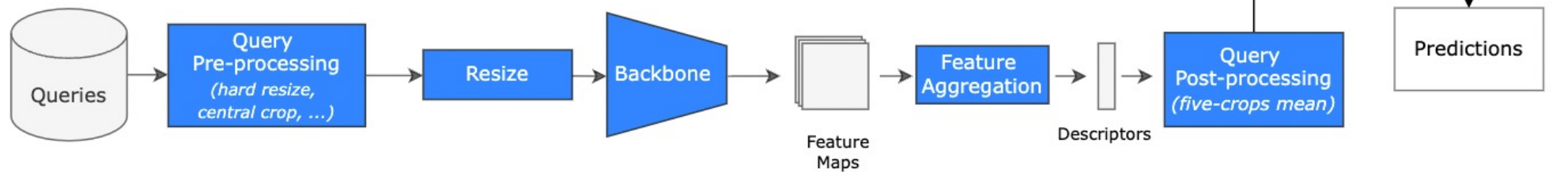


Test Time Scenario

Offline

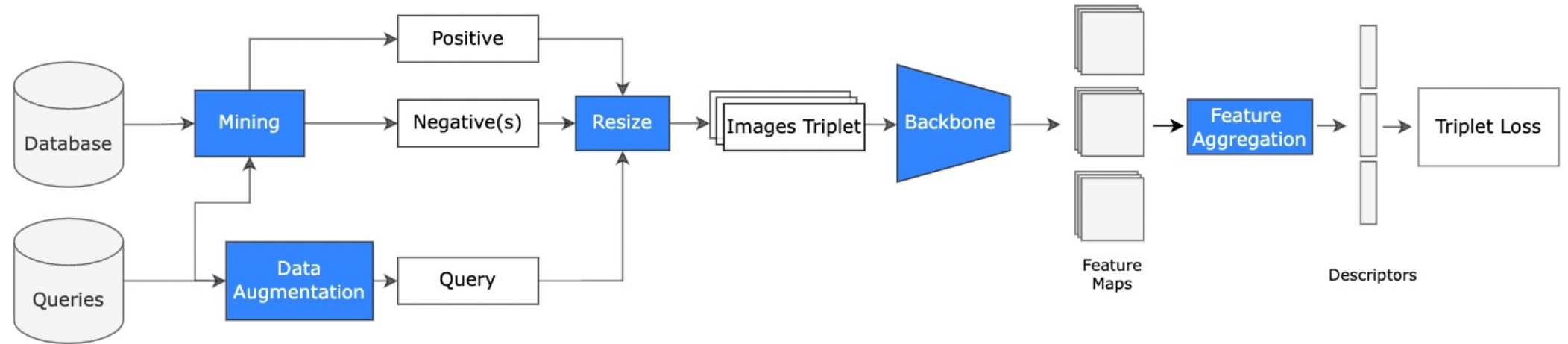


Online



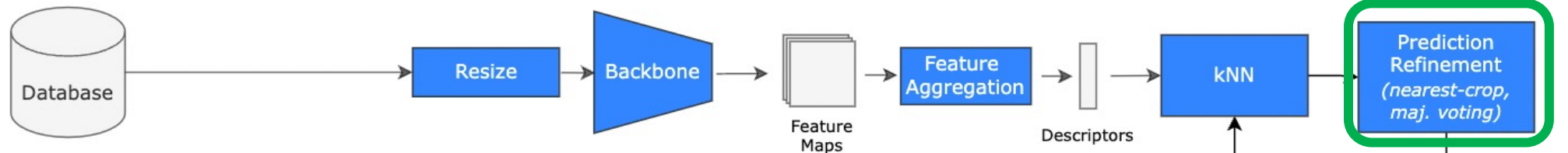
A visual geo-localization pipeline

Train Time Scenario

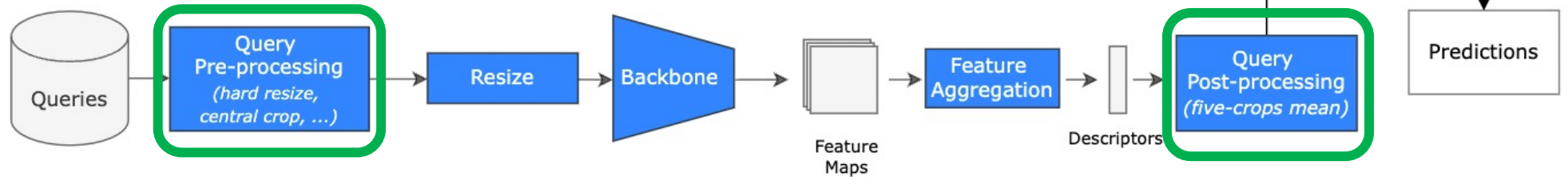


Test Time Scenario

Offline



Online



Metrics and parameters considered

- **Recall@N**: percentage of queries with one of the first N predictions within **25 meters**
- **Practical constraints**: relevant in a real-world scenario (GPU & CPU memory)
- **Computational cost**: FLOPs operations, model size, latency





Datasets

- In urban settings, the data can be systematically connected from cars equipped with onboard cameras. All the images in the database are from a **street level**.
- Aside from the type of sensor, the collected data may vary significantly depending on their position, on the acquisition campaign and on how it is structured.

6 Highly Diverse Datasets



Pitts 30k – 52k #images



MSLS – 1.5M #images



Tokyo 24/7 – 75k #images



R-SF – 1.0M #images



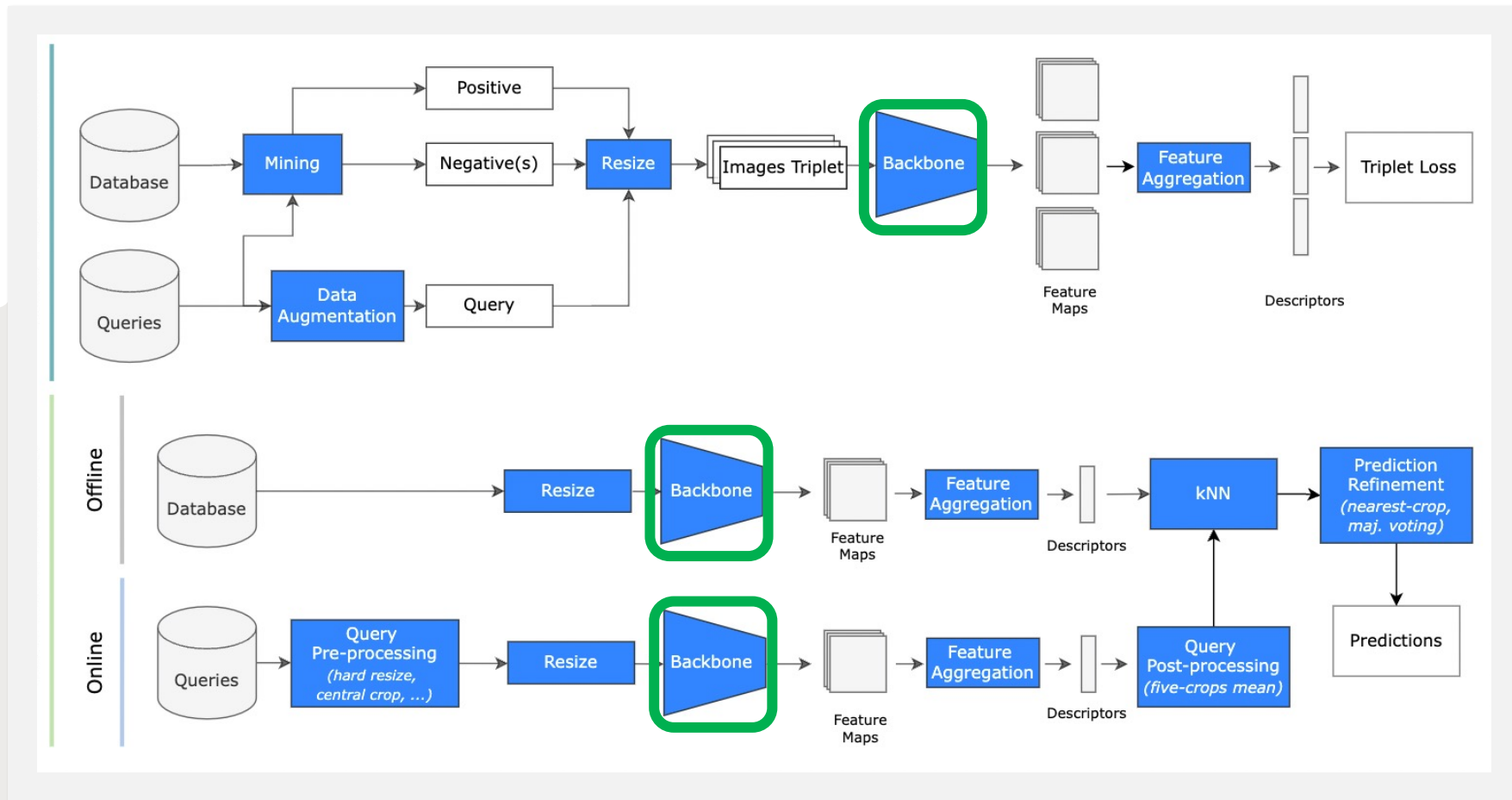
Eynsham – 48k #images



St. Lucia – 3k #images

- [1] Pitts30k, Arandjelovic et al, TPAMI 2018
- [2] MSLS, Warburg et al, CVPR 2020
- [3] Tokyo 24/7, Torii et al, TPAMI 2018

- [4] R-SF, Chen et al, CVPR 2011
- [5] Eynsham, Cummins and Newman, RSS 2009
- [6] St Lucia, Milford and Wyeth, T-RO 2008



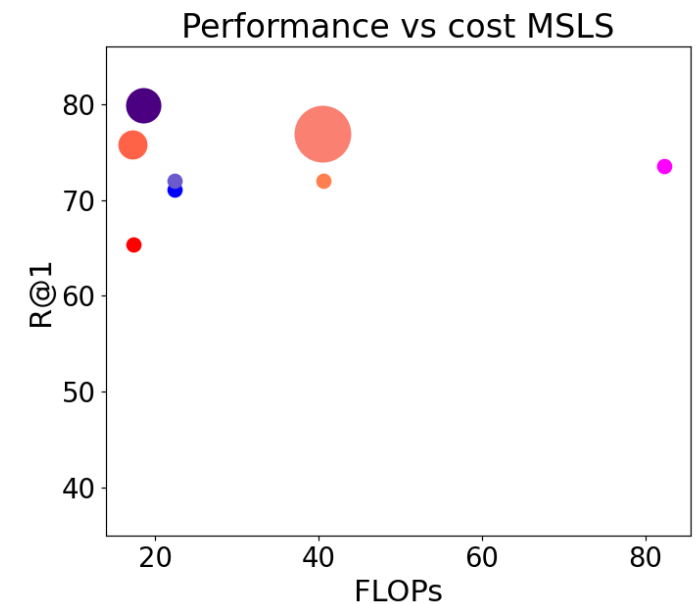
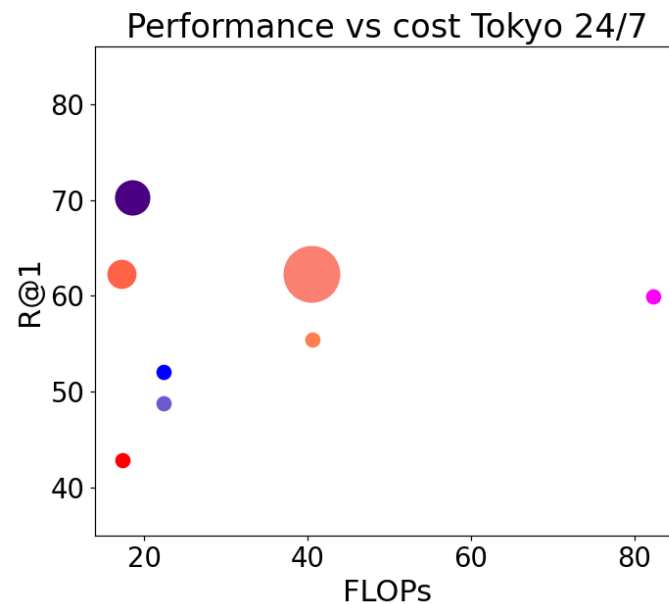
Backbones

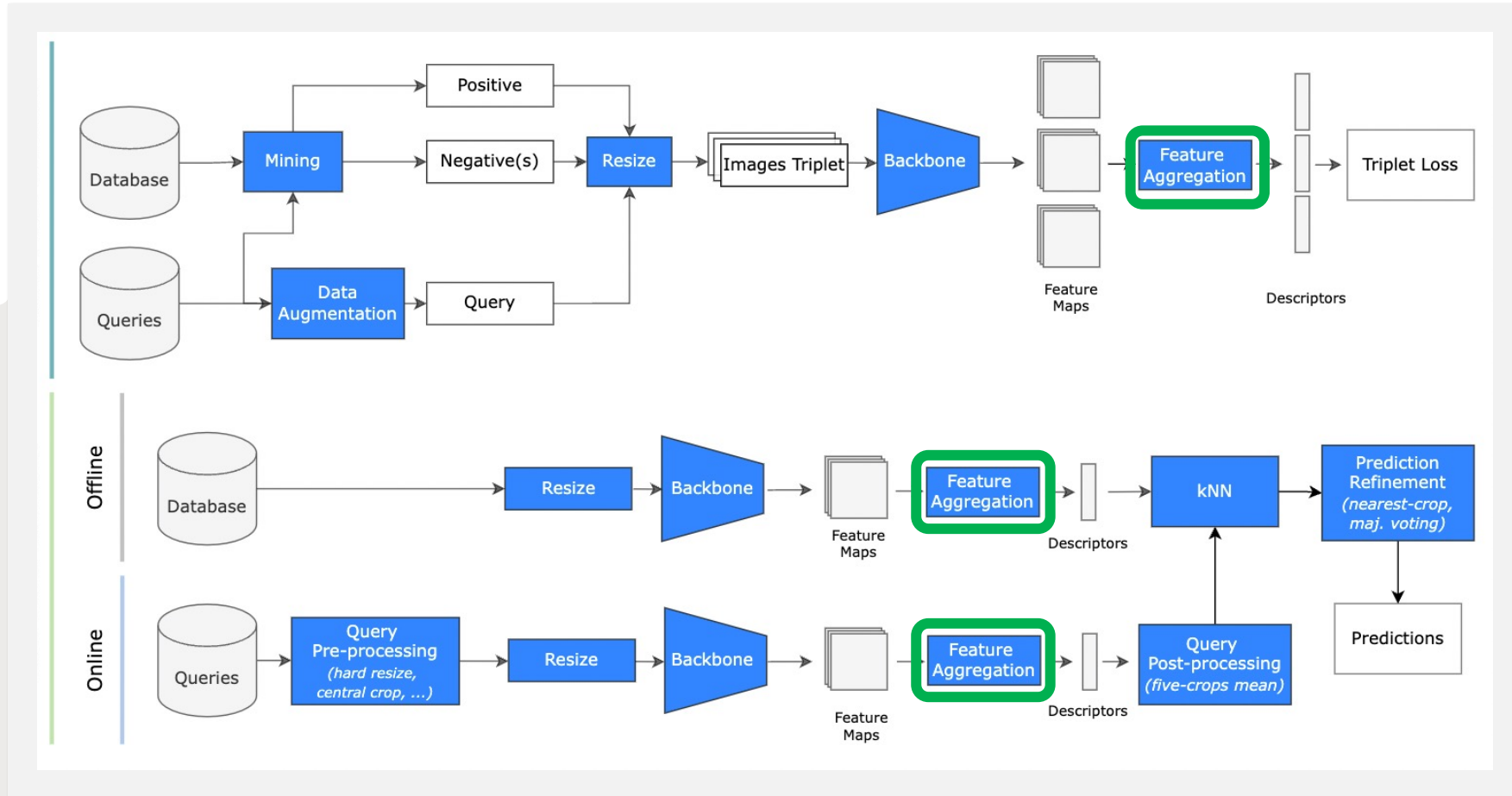
How does the choice of backbone affect results?

Tip n°1: backbones

- **ResNet-50** is an excellent choice as a CNN backbone, yielding close to the best results at a reasonable cost.
- **Compact Convolutional Transformers (CCT)** perform competitively well while being lightweight.
- Different backbones may perform better if **cropped before the last layer** (e.g., ResNets cropped at the *conv_4* layer).

● ViT + CLS	d=768
● CCT + CLS	d=384
● CCT + GeM	d=384
● CCT + NetVLAD	d=24576
● ResNet-18 + GeM	d=256
● ResNet-50 + GeM	d=1024
● ResNet-18 + NetVLAD	d=16384
● ResNet-50 + NetVLAD	d=65536



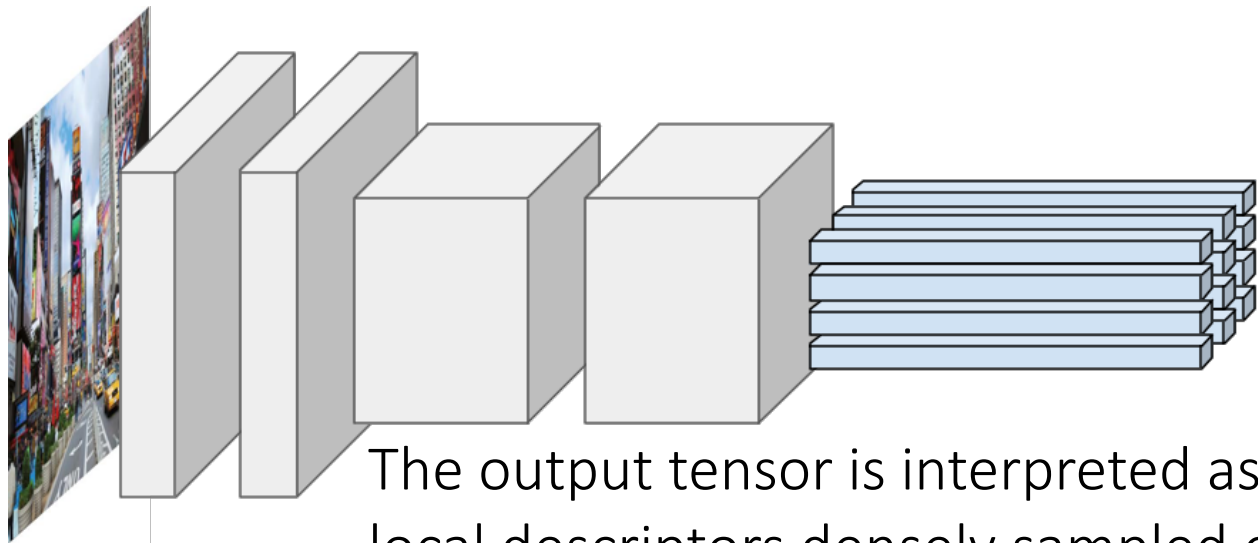


Aggregation layers

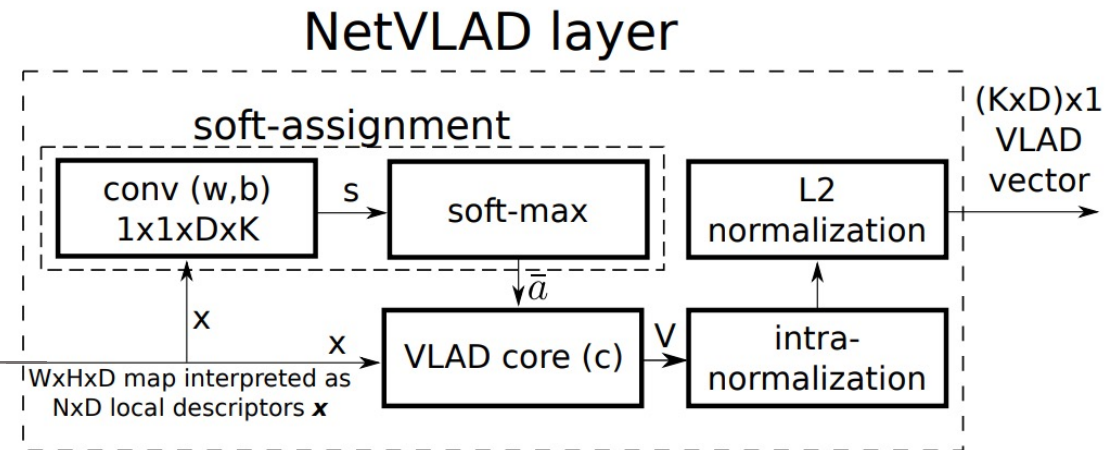
To cluster or to pool? This is the dilemma

NetVLAD

The NetVLAD layer clusters local descriptors via a soft-assignment w.r.t. a dictionary of visual words. The residuals of the local descriptors w.r.t. to the center of the clusters are summed to form a single Global Descriptor

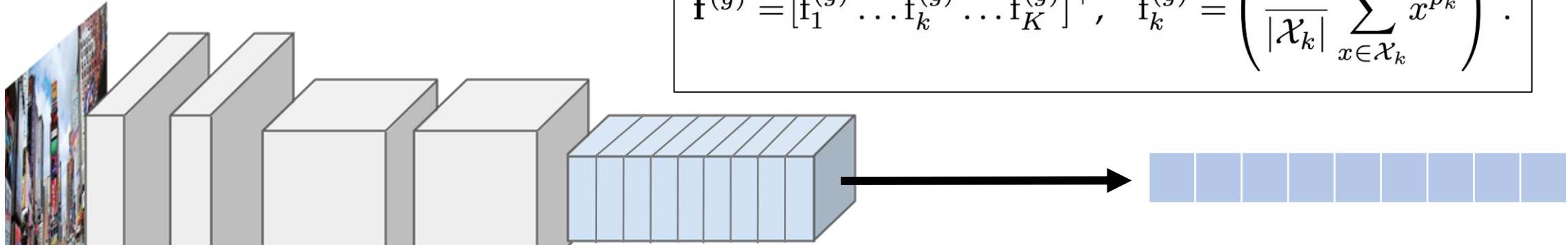


The output tensor is interpreted as a set of $H \times W$ local descriptors densely sampled over the input image (i.e., without keypoint detection)



GeM

The **Generalized Mean** layer summarizes the content at each output feature map with a pooling operation that generalizes max and mean pooling.



$$\mathbf{f}^{(g)} = [f_1^{(g)} \dots f_k^{(g)} \dots f_K^{(g)}]^\top, \quad f_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k} \right)^{\frac{1}{p_k}}$$

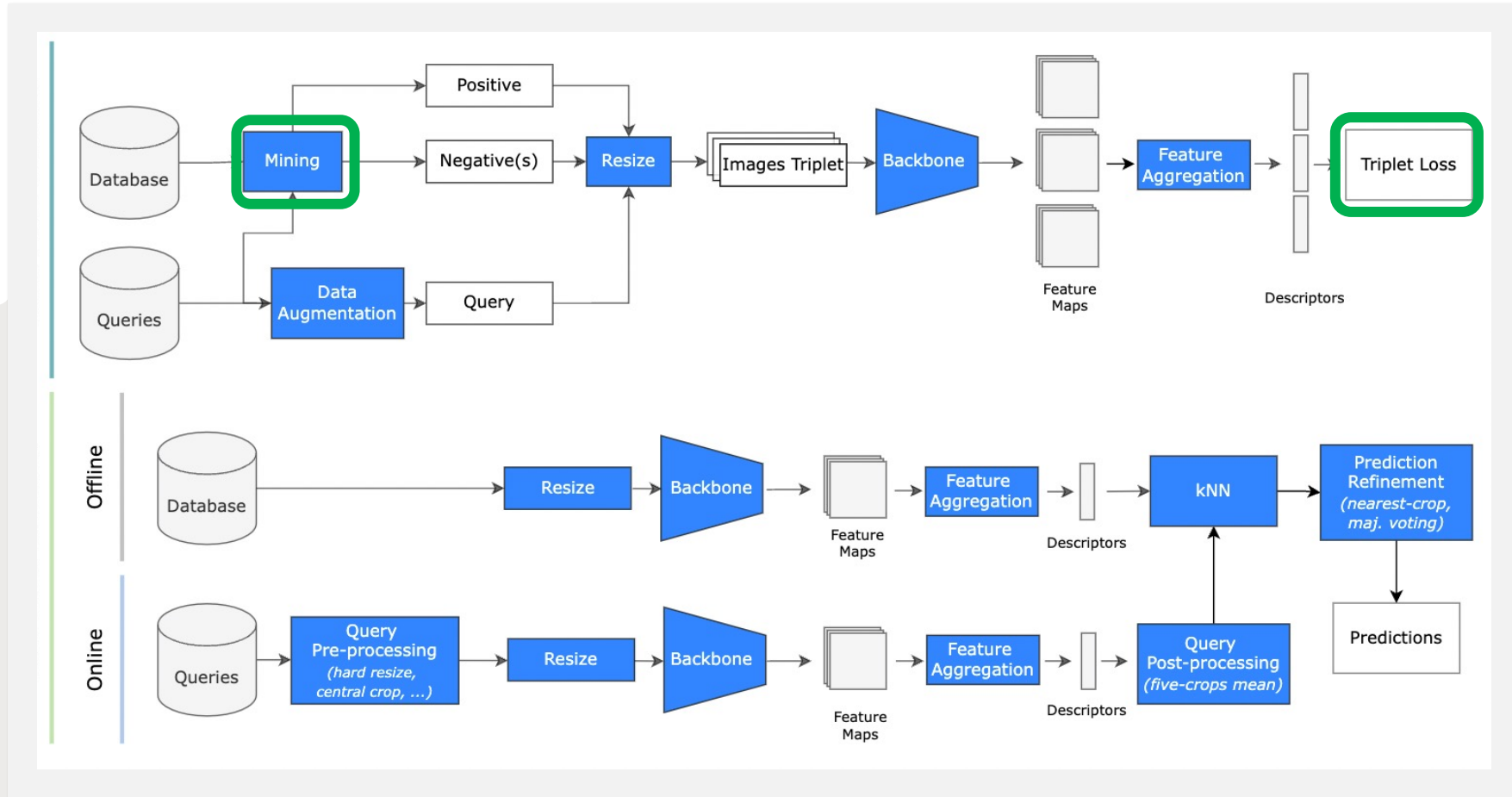
The output D feature maps are summarized to a D-dimensional vector

Tip n°2: aggregation

Clustering based methods (NetVLAD, CRN) lead to **better performance**.

GeM pooling, which is much more efficient, has shown a **better generalization power**, especially when training the model on a large and heterogeneous dataset

Backbone	Aggregation Method	Features Dim	Training on Pitts30k						Training on MSLS						
			R@1 Pitts30k	R@1 MSLS	R@1 Tokyo 24/7	R@1 R-SF	R@1 Eynsham	R@1 St Lucia	R@1 Pitts30k	R@1 MSLS	R@1 Tokyo 24/7	R@1 R-SF	R@1 Eynsham	R@1 St Lucia	R@1 Average
ResNet-50	GeM	1024	82.0	38.0	41.5	45.4	66.3	59.0	77.4	72.0	55.4	45.7	83.9	91.2	63.2
ResNet-50	NetVLAD + PCA 1024	1024	83.9	46.5	59.4	53.2	72.5	57.7	77.4	74.8	51.3	39.0	85.2	92.9	66.2
ResNet-50	CRN + PCA 1024	1024	84.1	49.9	64.6	58.8	74.3	63.4	77.3	75.6	51.8	38.8	85.7	94.1	68.2
ResNet-50	GeM + FC 2048	2048	80.1	33.7	43.6	48.2	70.0	56.0	79.2	73.5	64.0	55.1	86.1	90.3	65.0
ResNet-50	NetVLAD + PCA 2048	2048	84.4	47.9	62.6	56.0	74.1	58.9	78.5	75.4	52.8	42.6	85.8	93.4	67.7
ResNet-50	CRN + PCA 2048	2048	84.7	51.2	67.1	62.3	75.8	65.0	78.3	76.3	54.3	42.8	86.2	94.4	69.9
ResNet-50	GeM + FC 65536	65536	80.8	35.8	45.6	49.0	72.5	59.6	79.0	74.4	69.2	58.4	86.2	90.8	66.8
ResNet-50	NetVLAD	65536	86.0	50.7	69.8	67.1	77.7	60.2	80.9	76.9	62.8	51.5	87.2	93.8	72.1
ResNet-50	CRN	65536	85.8	54.0	73.1	70.9	79.7	65.9	80.8	77.8	63.6	53.4	87.5	94.8	73.9

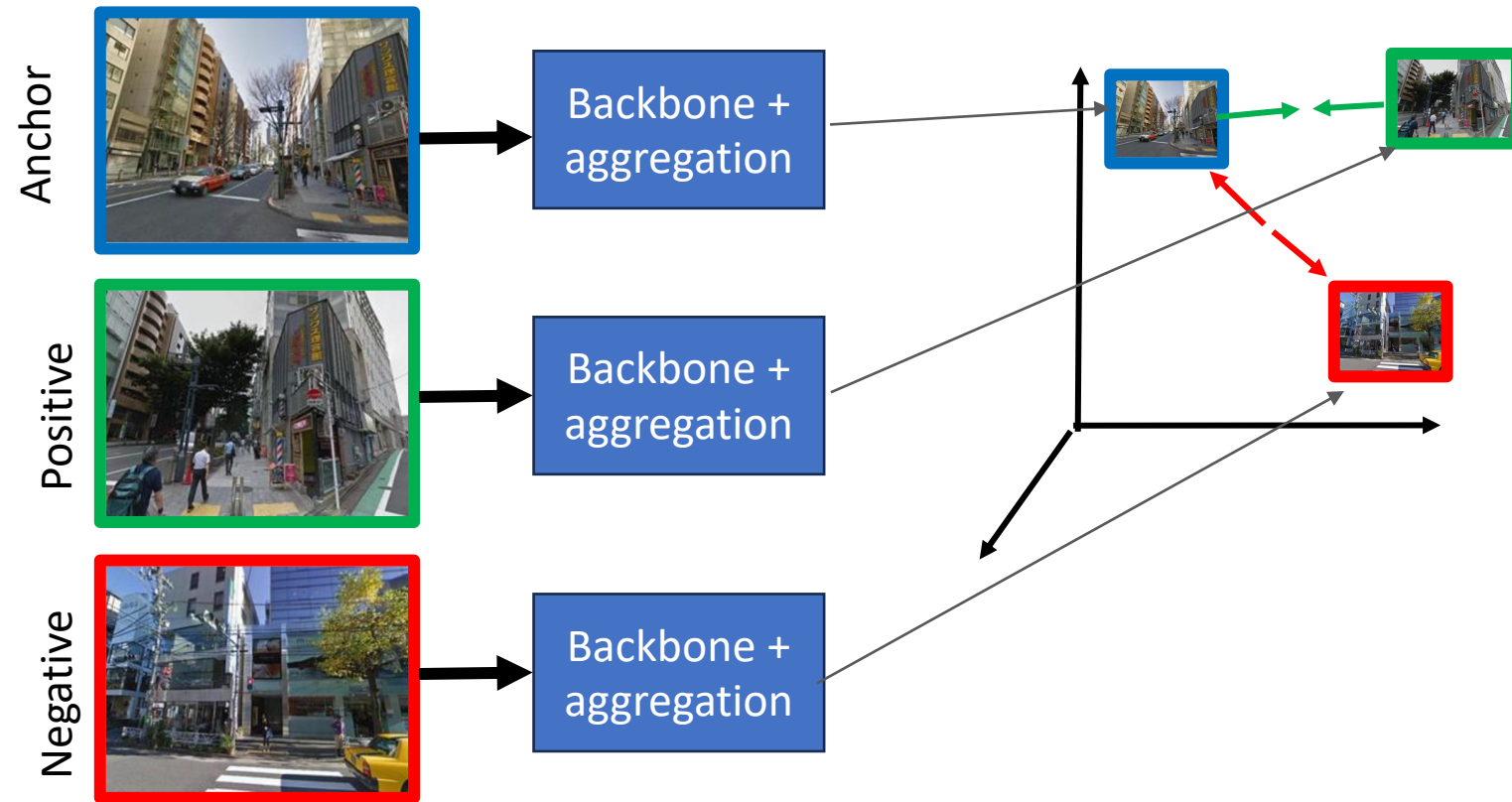


Contrastive learning and mining

Is full mining necessary?

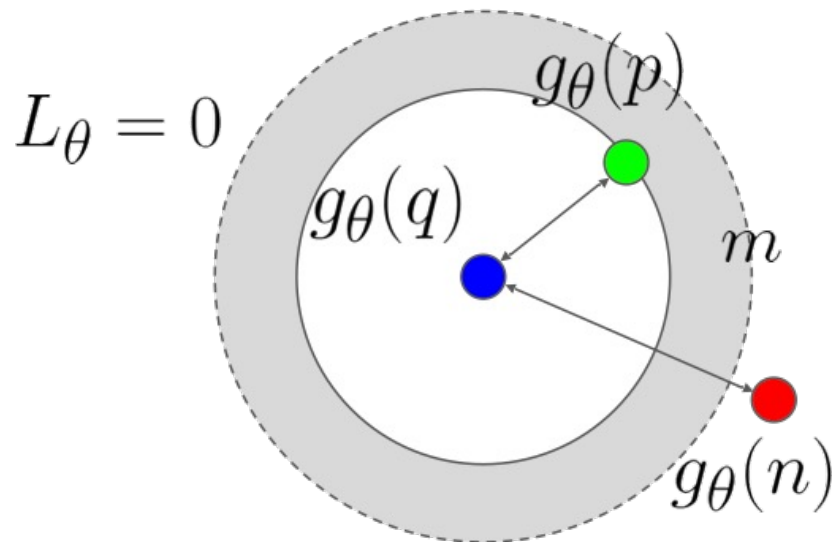
Contrastive Learning

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



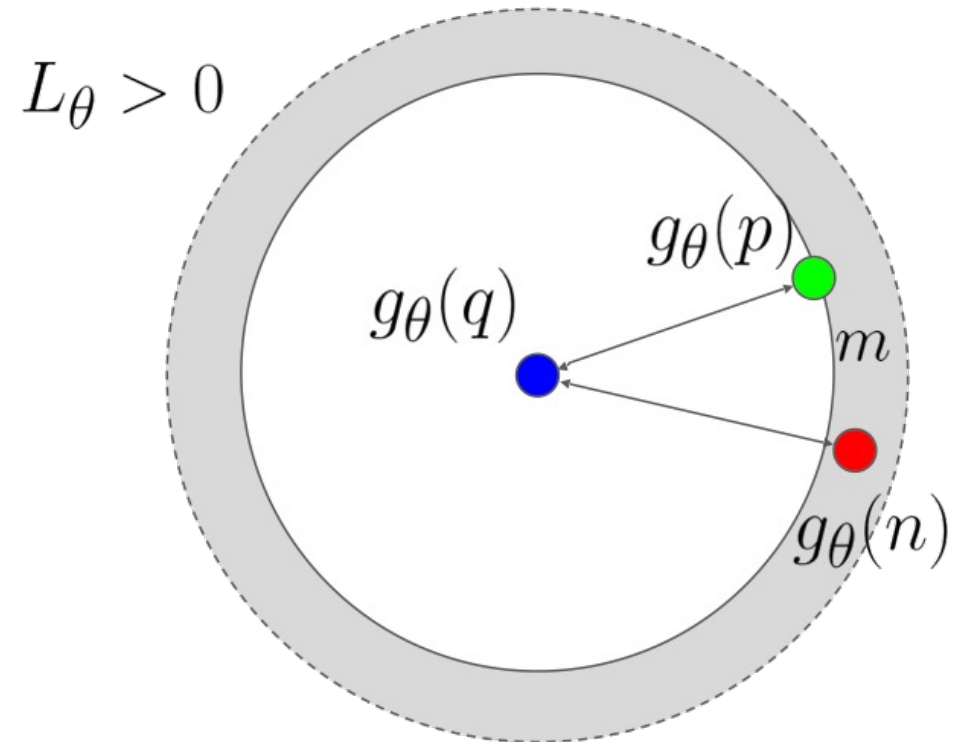
Mining

Poor choice of positive/negative examples: they already satisfy the margin



The choice of positives and negatives affects training, and depends on the representation learned by the model, so it must be repeated during training.

Good choice of positive/negative examples: the margin is not satisfied



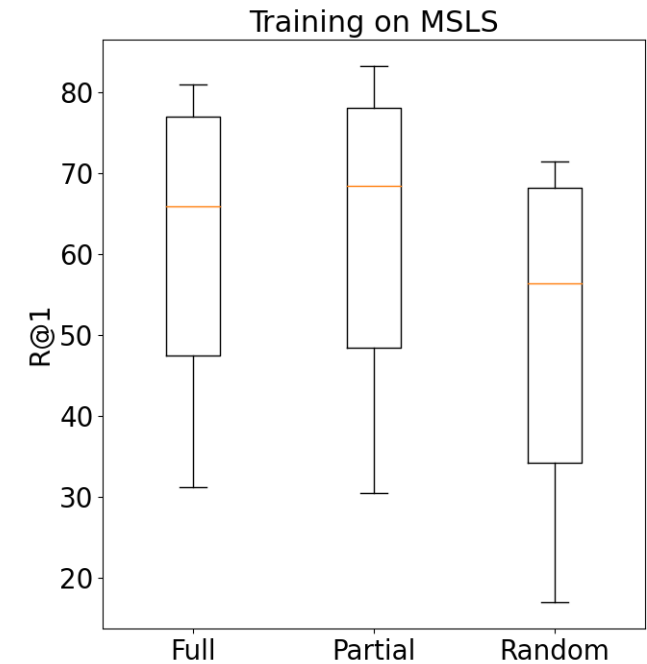
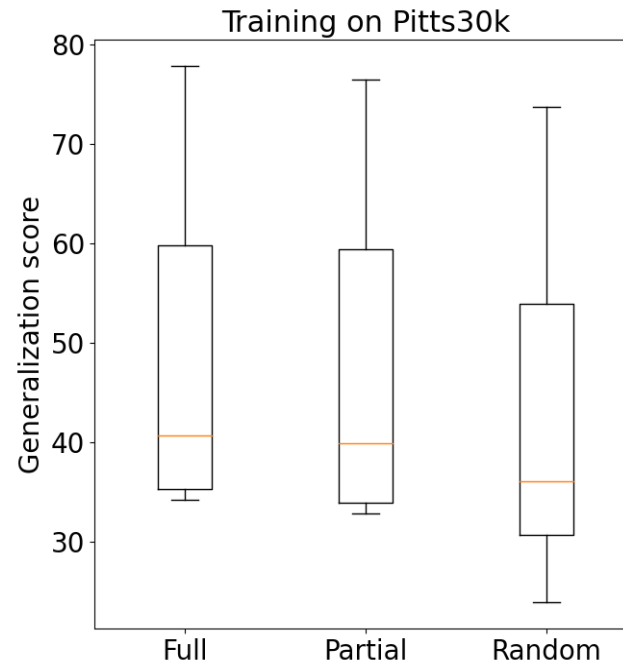
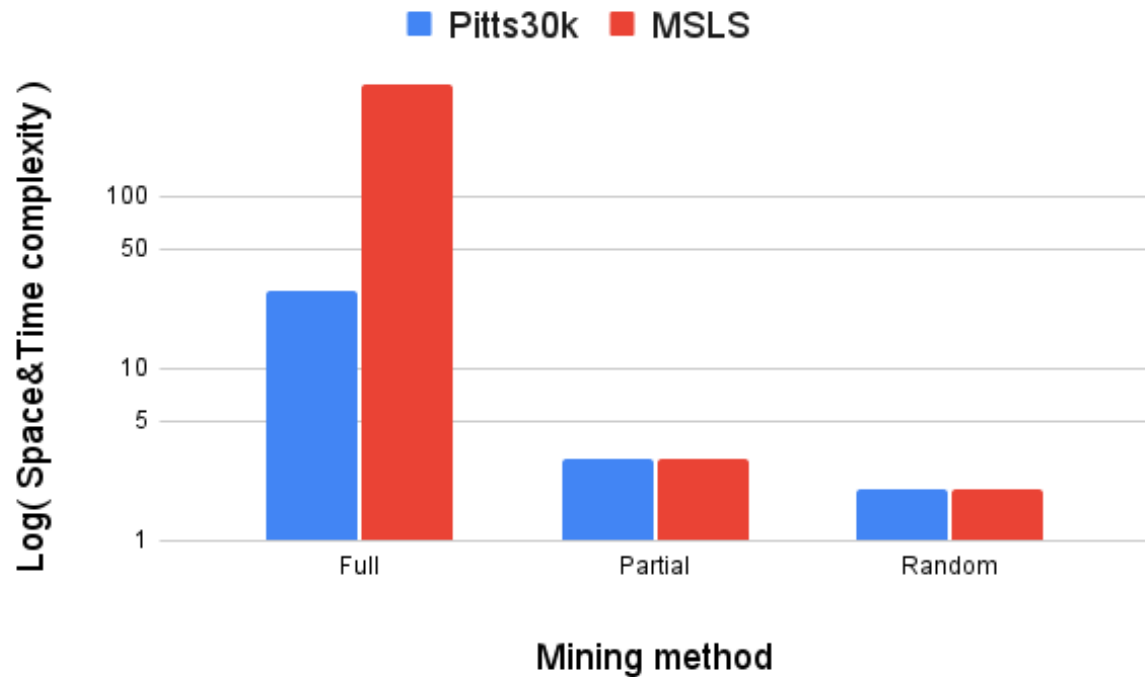
Tip n°3: mining

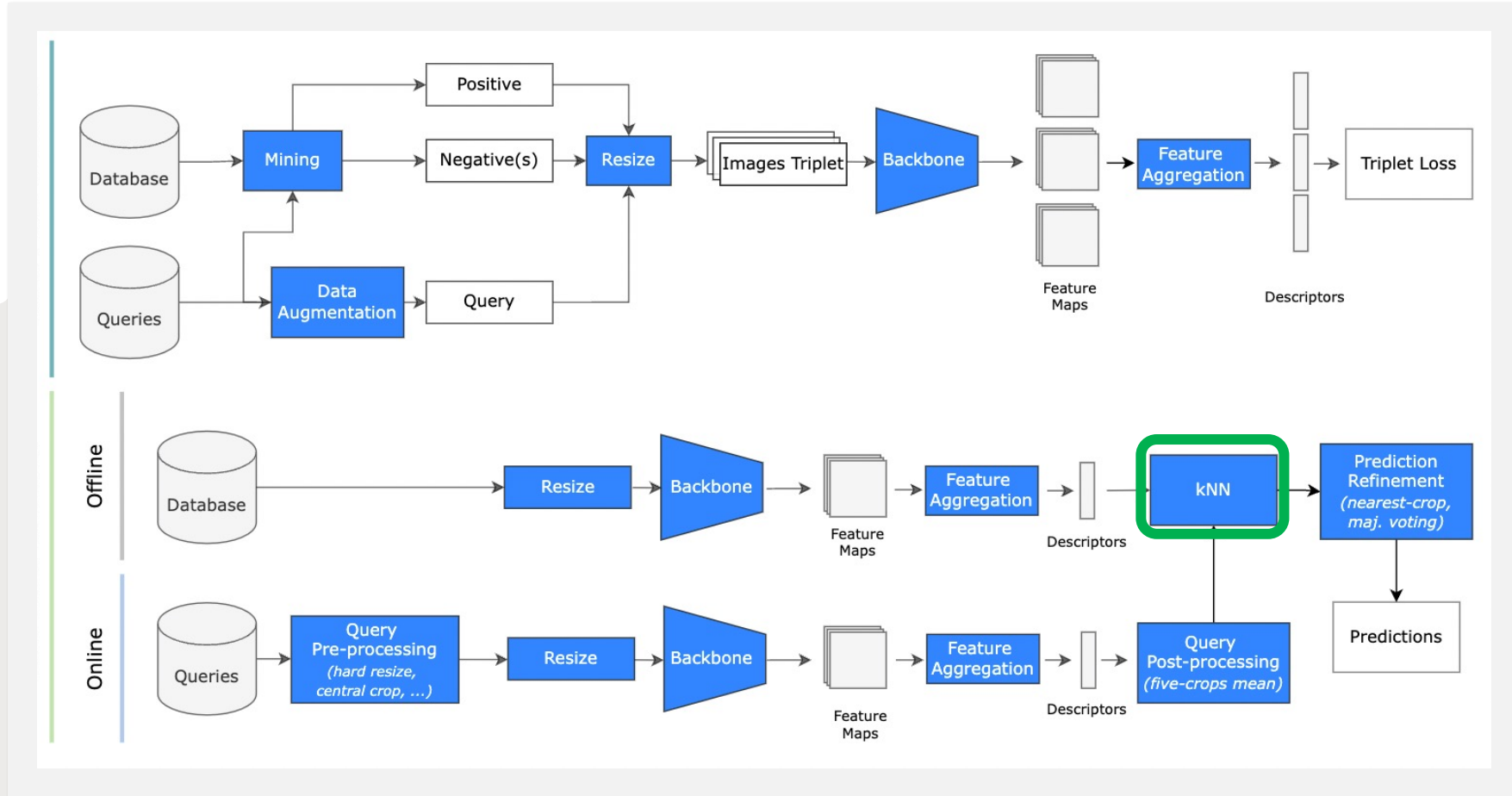
- Partial mining yields similar or sometimes even higher results than full mining at a fraction of the cost.
- On large databases, full mining is not feasible in a reasonable time

Backbone	Aggregation Method	Mining Method	Space & Time Complexity	Training on Pitts30k			Training on MSLS		
				R@1 Pitts30k	R@1 MSLS	R@1 Tokyo 24/7	R@1 Pitts30k	R@1 MSLS	R@1 Tokyo 24/7
ResNet-18	GeM	Random	$\mathcal{O}(1)$	73.7	30.5	31.3	62.2	50.6	28.8
ResNet-18	GeM	Full database	$\mathcal{O}(\#db + \#q)$	77.8	35.3	35.3	70.1	61.8	42.8
ResNet-18	GeM	Partial database	$\mathcal{O}(k_{db} + k_q + \#pos)$	76.5	34.2	33.9	71.6	65.3	42.8
ResNet-18	NetVLAD	Random	$\mathcal{O}(1)$	83.9	43.6	55.1	73.3	61.5	45.0
ResNet-18	NetVLAD	Full database	$\mathcal{O}(\#db + \#q)$	86.4	47.4	63.4	-	-	-
ResNet-18	NetVLAD	Partial database	$\mathcal{O}(k_{db} + k_q + \#pos)$	86.2	47.3	61.2	81.6	75.8	62.3

Tip n°3: mining

- Partial mining yields similar or sometimes even higher results than full mining at a fraction of the cost.
- On large databases, full mining is not feasible in a reasonable time



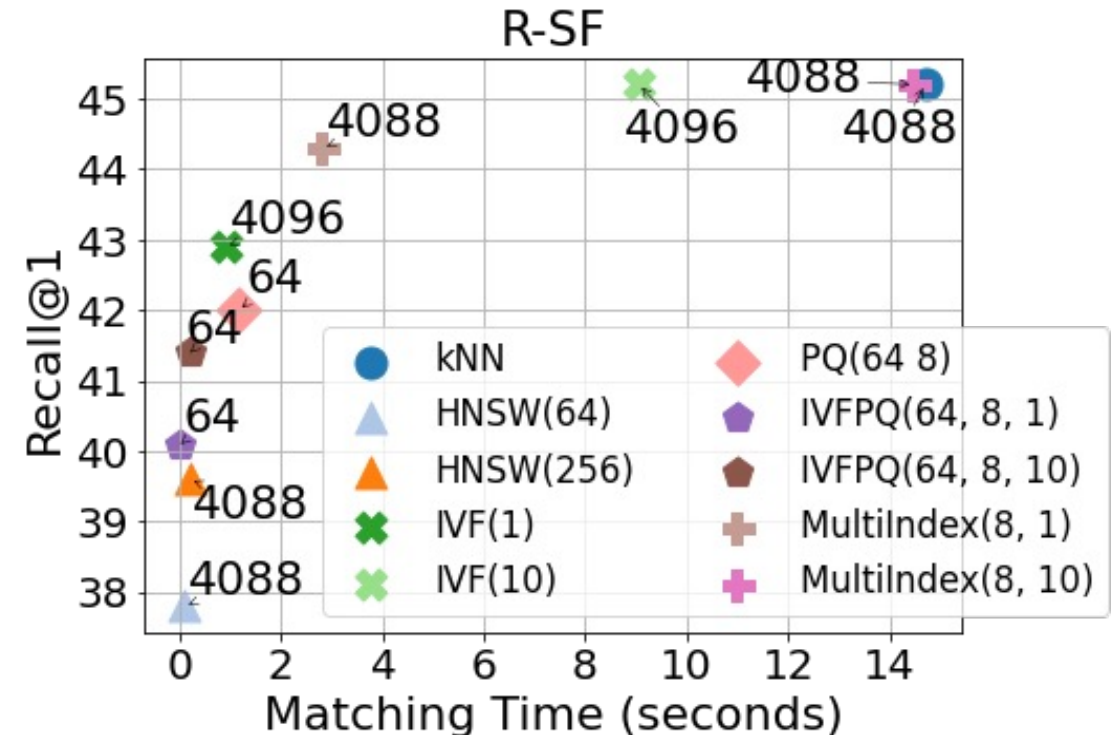
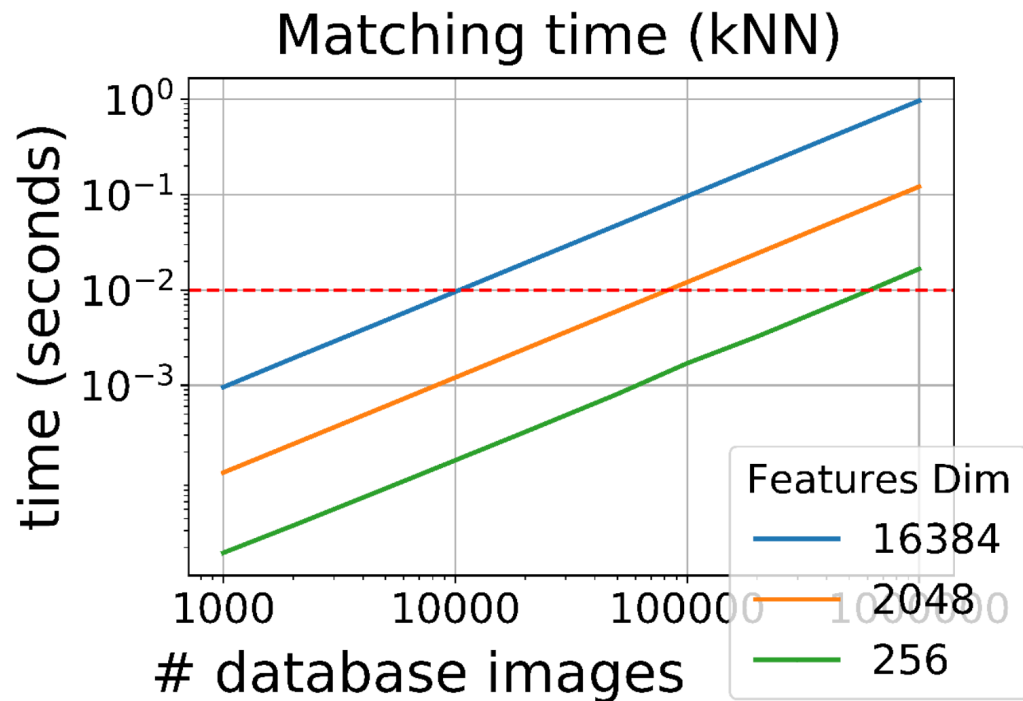


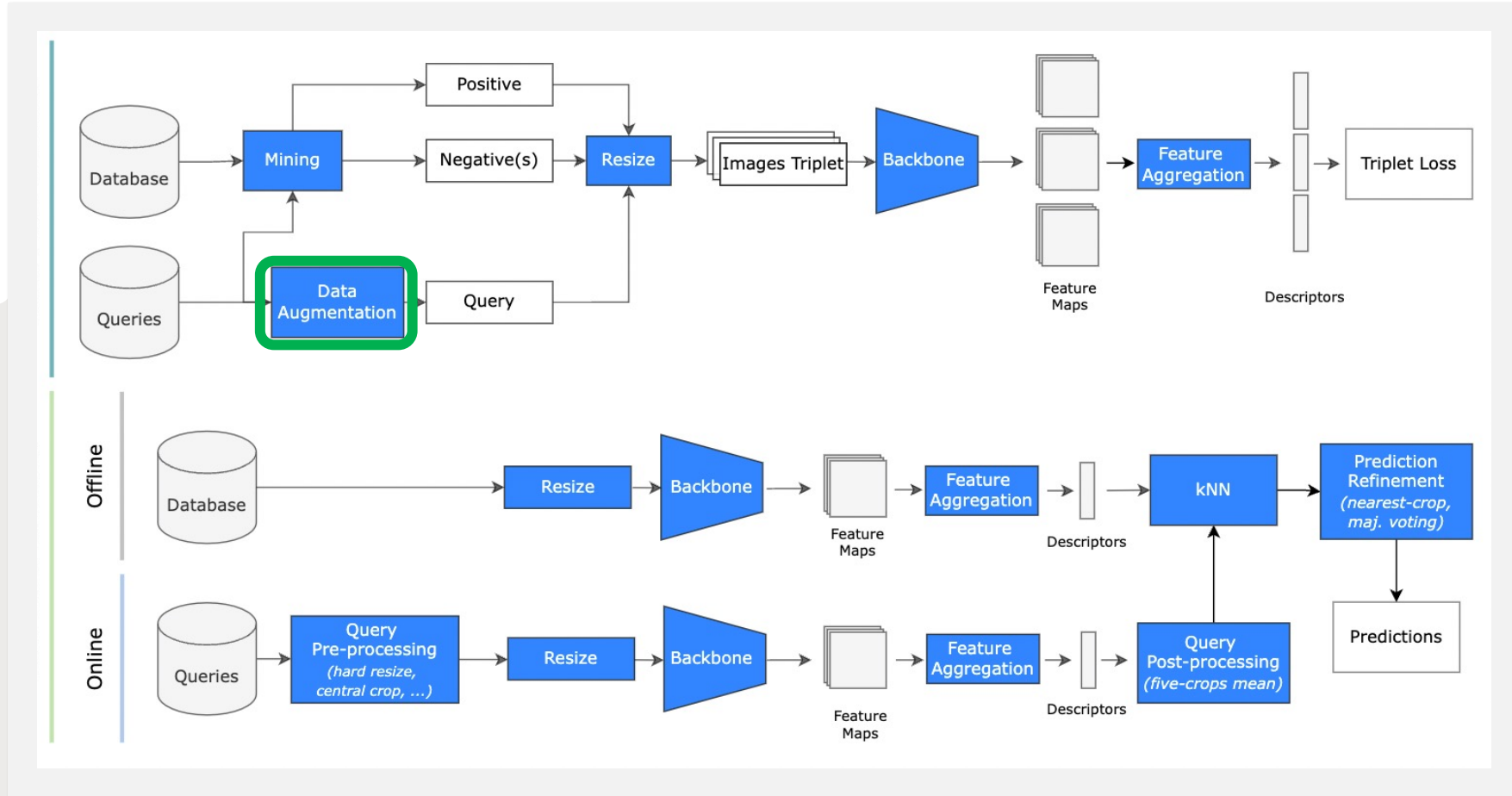
Nearest neighbor search

Memory and time constraints!

Tip n°4: kNN

- **Product quantization** with an **inverted file index** can lead to a drastic reduction of the matching time and memory requirement, even up to 98.5%, only with a small performance drop.
- The **inverted multi-index** provides an 80% saving in matching time, losing only a 0.9 % of recall.



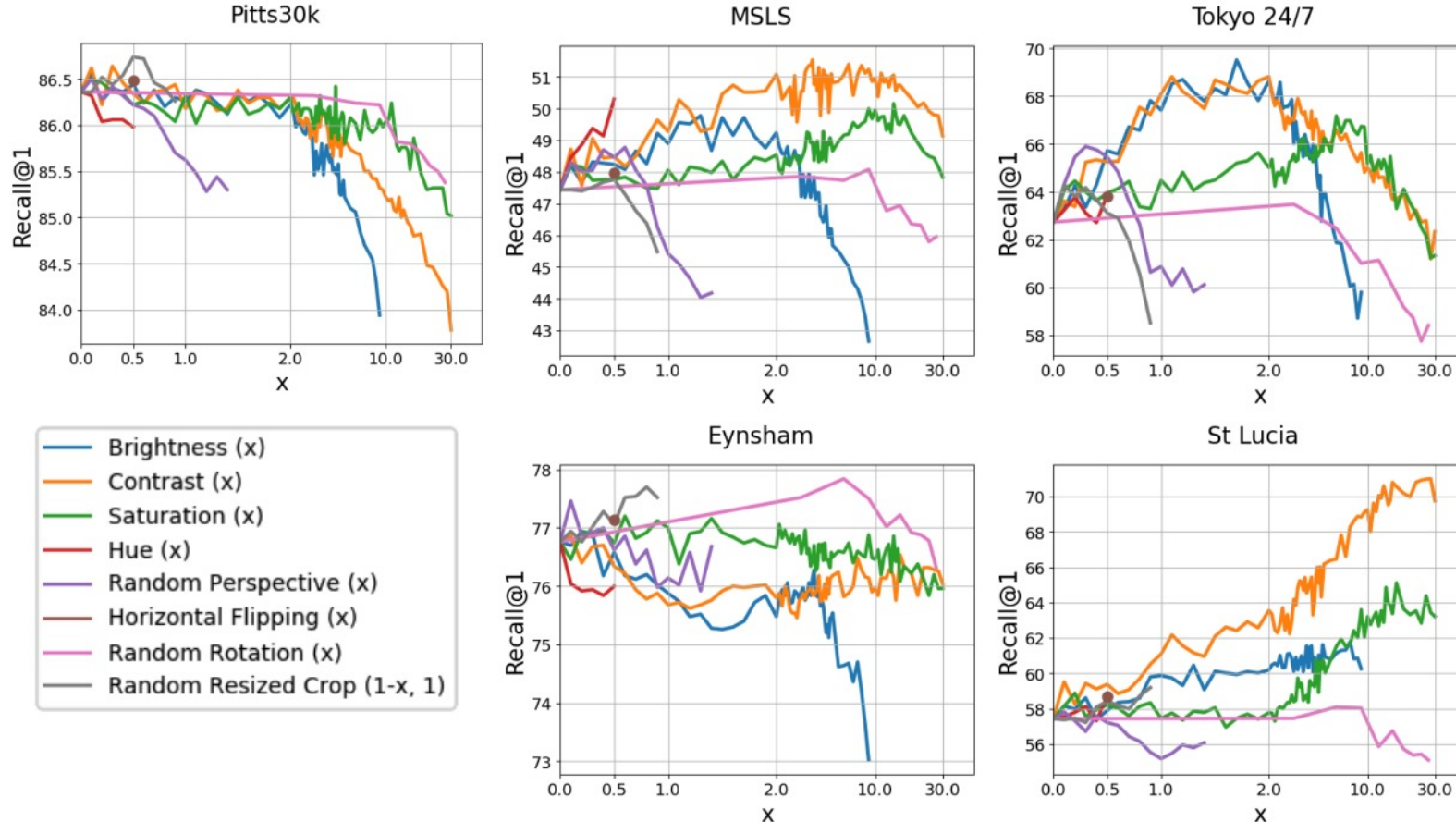


Data augmentation

Does one data augmentation fit all datasets?

Tip n°5: data augmentation

- The effectiveness of the color jittering augmentations are highly dependent on the dataset.
- Horizontal flipping and resized cropping provide a slight but consistent boost in all cases.



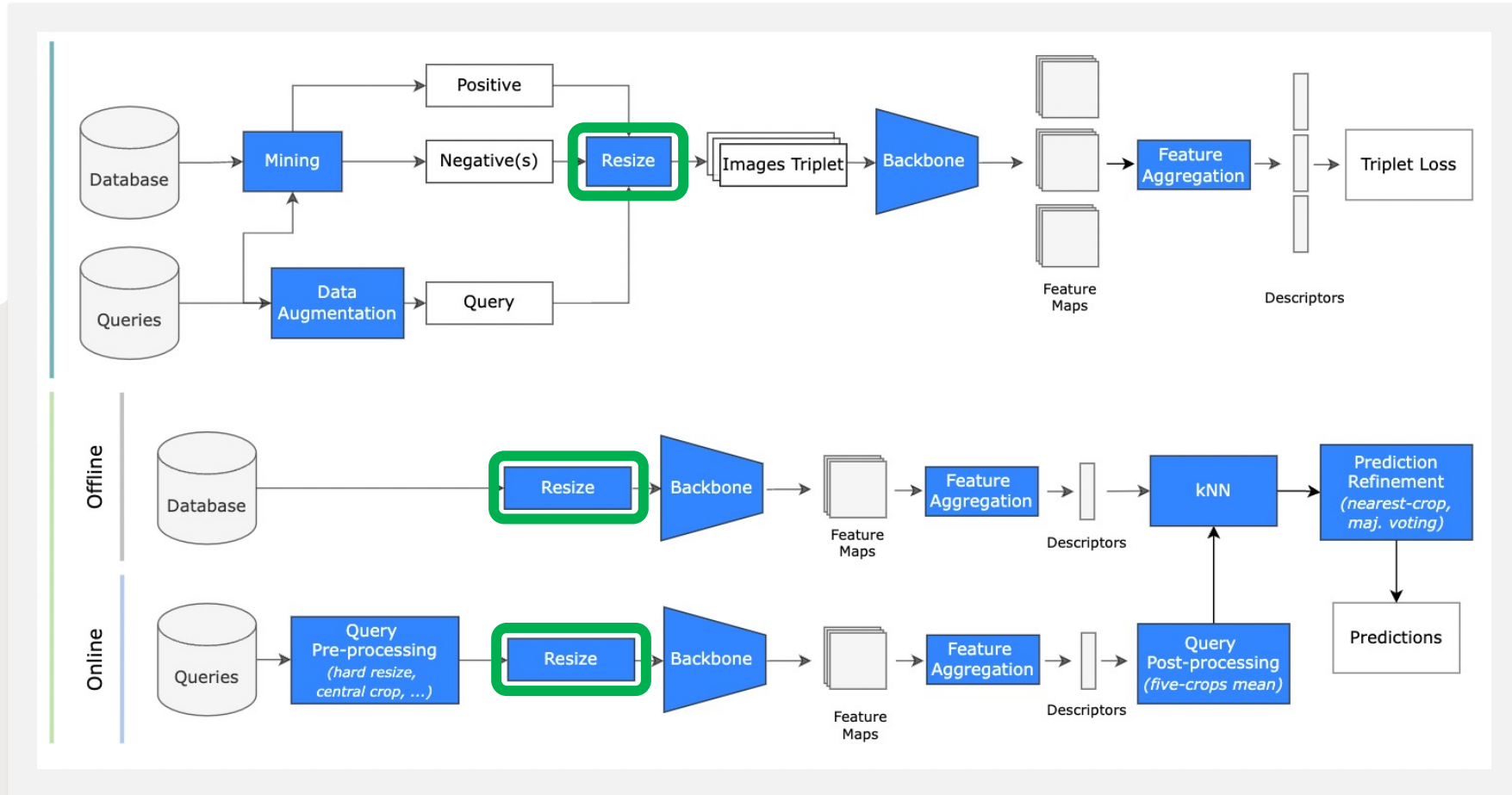
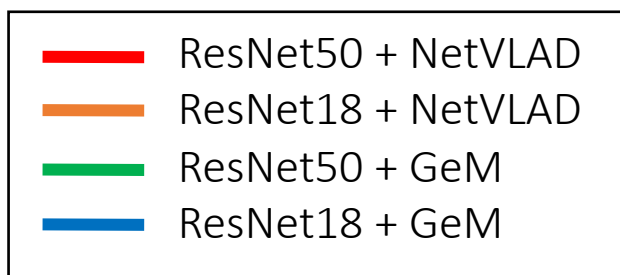
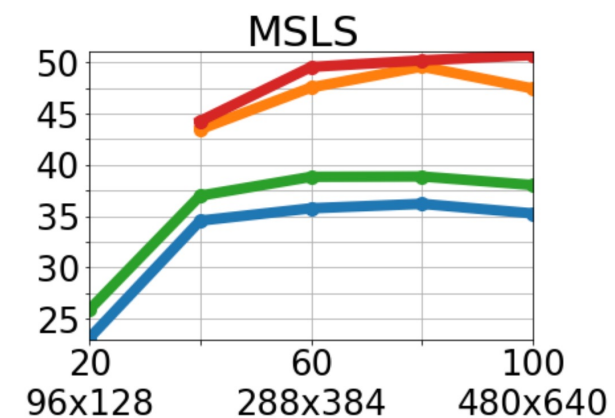
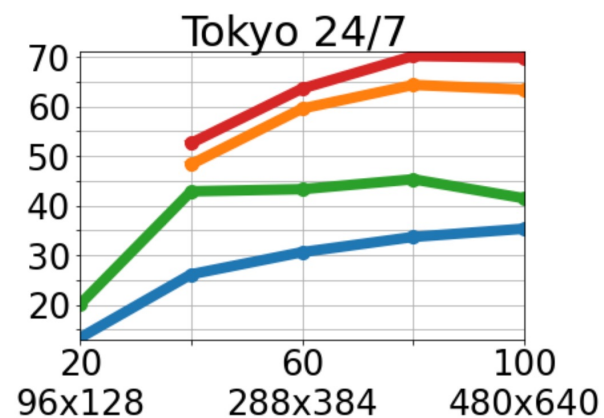
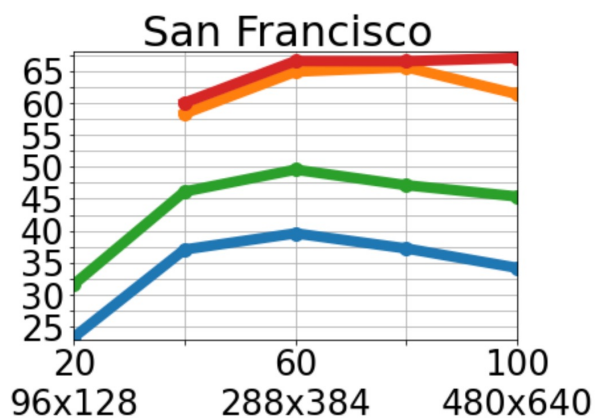
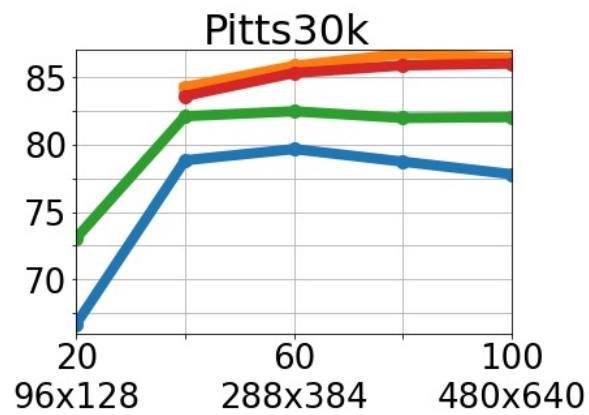


Image resolution

Is full resolution necessary for visual geo-localization?

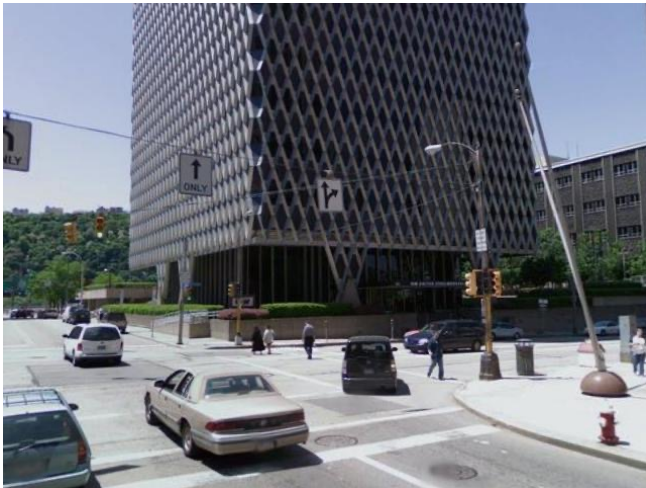
Tip n°6: image resolution

- NetVLAD's descriptors seem to better handle higher resolutions than their GeM counterpart
- Lower resolutions, as low as 40%, show improved results especially when there is a wide domain gap between train and test sets (e.g., Pitts30k -> St. Lucia).
- 40% resolution means reducing it FLOPs by 16% w.r.t. full resolution images. Storage needs also decrease!



Tip n°6: image resolution

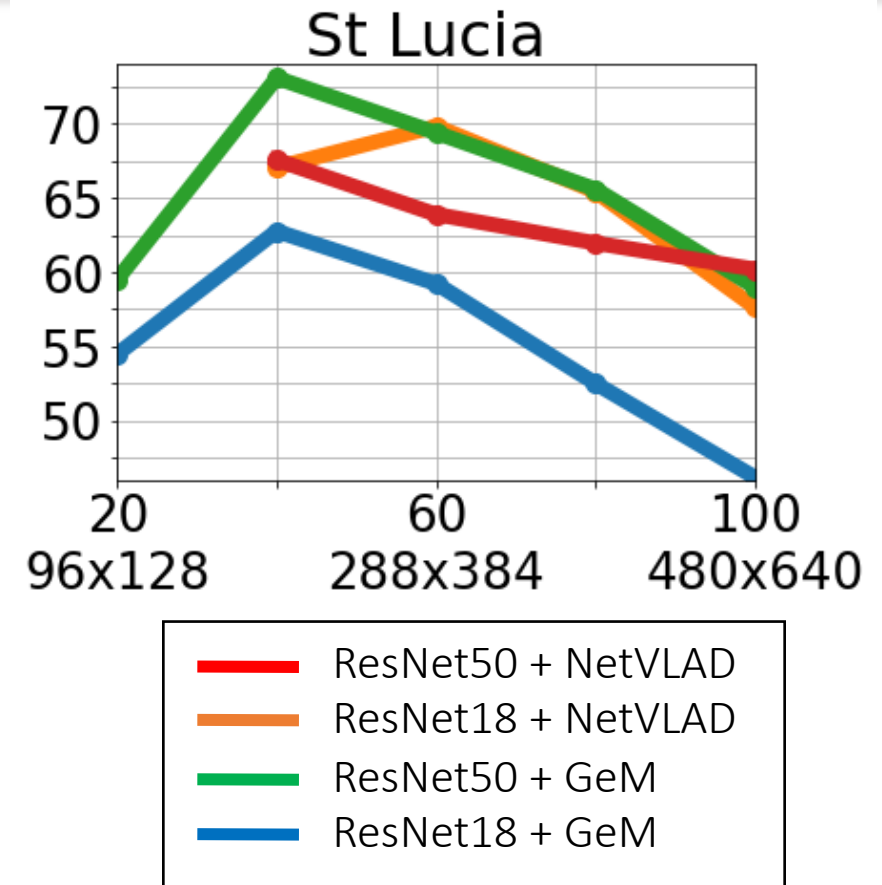
- NetVLAD's descriptors seem to better handle higher resolutions than their GeM counterpart
- Using the highest available resolution is in most cases superfluous. Lower resolutions, as low as 40%, show improved results especially when there is a wide domain gap between train and test sets (e.g., Pitts30k -> St. Lucia)



Pitts30k

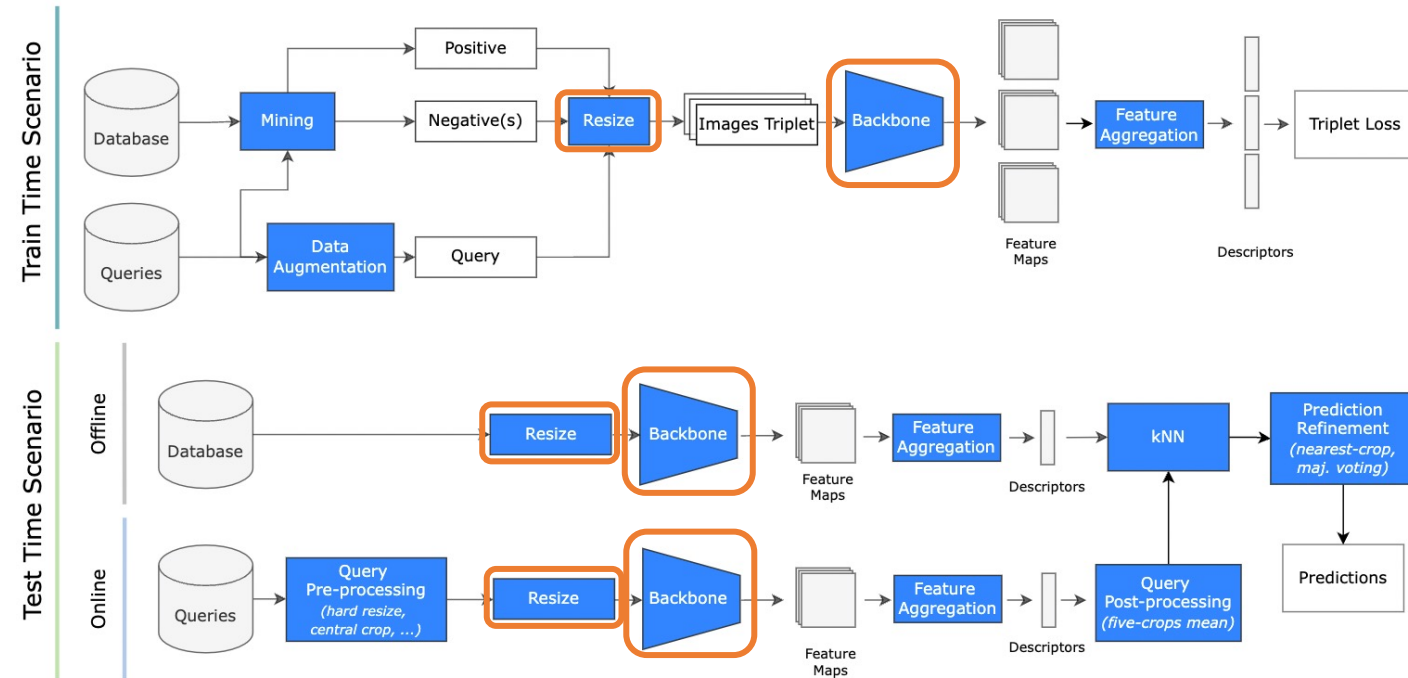


St Lucia



Using what we learned

Let's consider a pipeline based on NetVLAD.



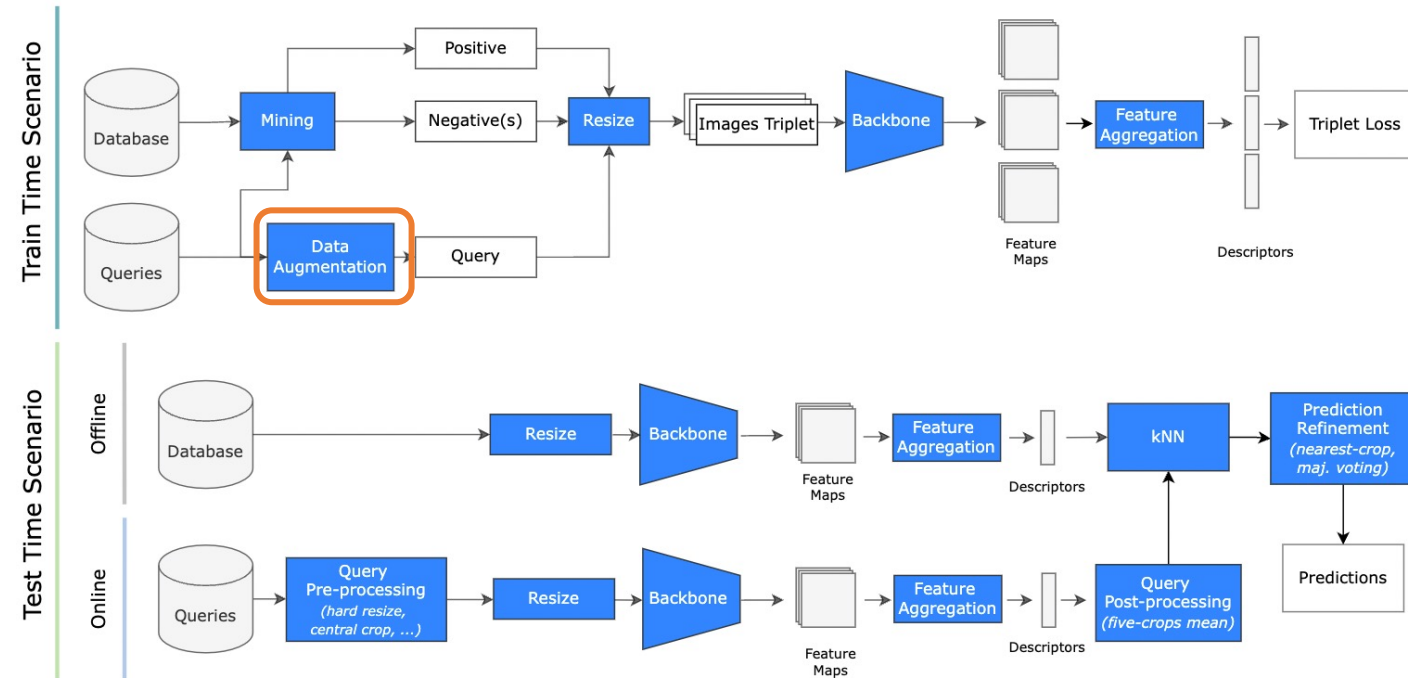
- Backbone: ResNet-18
- Resize: 80% resolution



faster extraction, lower memory requirements

Using what we learned

Let's consider a pipeline based on NetVLAD.



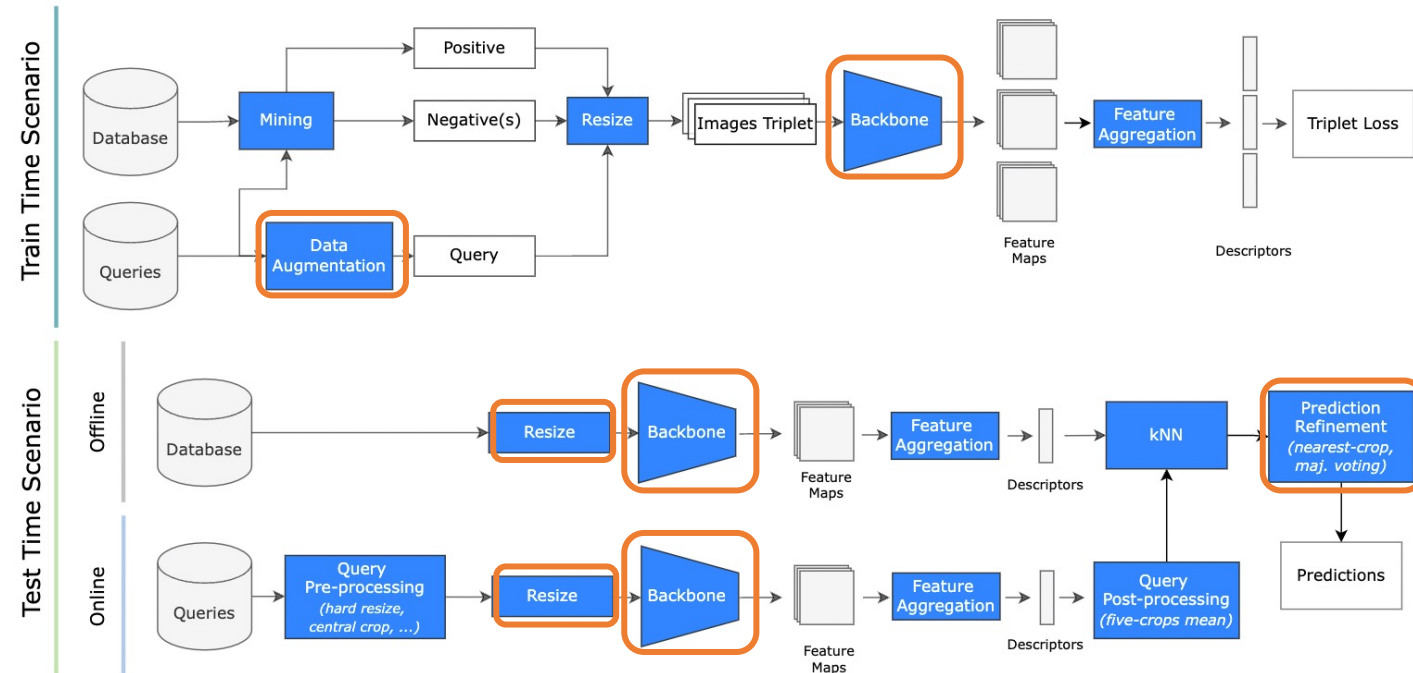
- Data Augmentation



increased robustness

Using what we learned

With some careful design choices, simple architectures can reach competitive results!



Method	Feat. Dim	R@1 Tokyo 24/7
SRALNet [1]	32768	68.6
APPSVR [2]	32768	68.3
ResNet18 + NetVLAD (ours)	16384	73.7

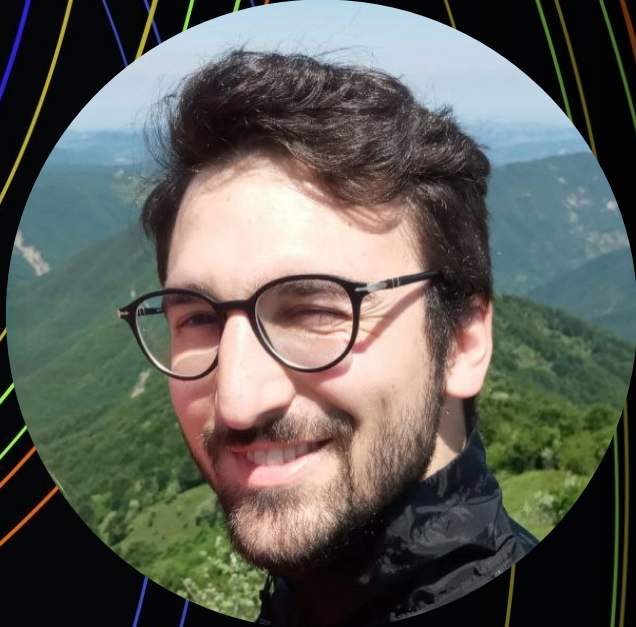
[1] G. Peng et al., "Semantic reinforced attention learning for visual place recognition", ICRA 2021

[2] G. Peng et al., "Attentional pyramid pooling of salient visual residuals for place recognition", ICCV 2021

Code and models

Code of the benchmark
and trained models
are available at this link





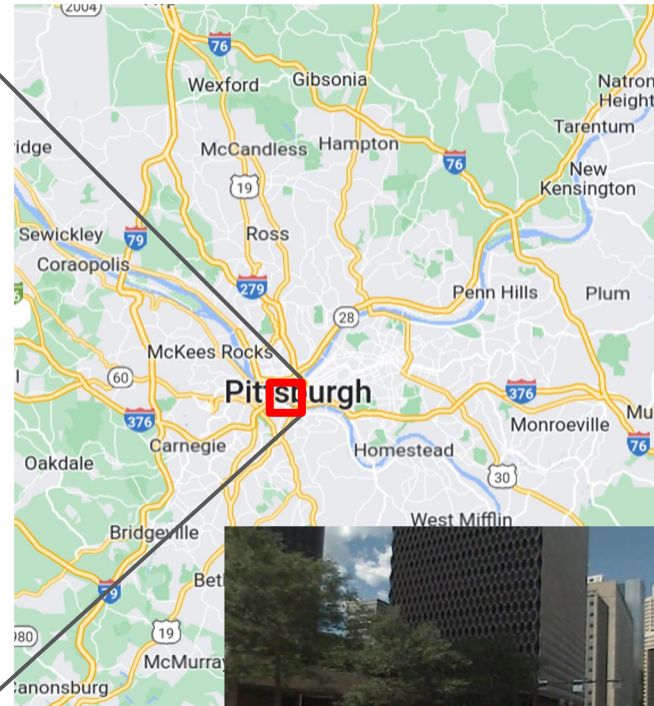
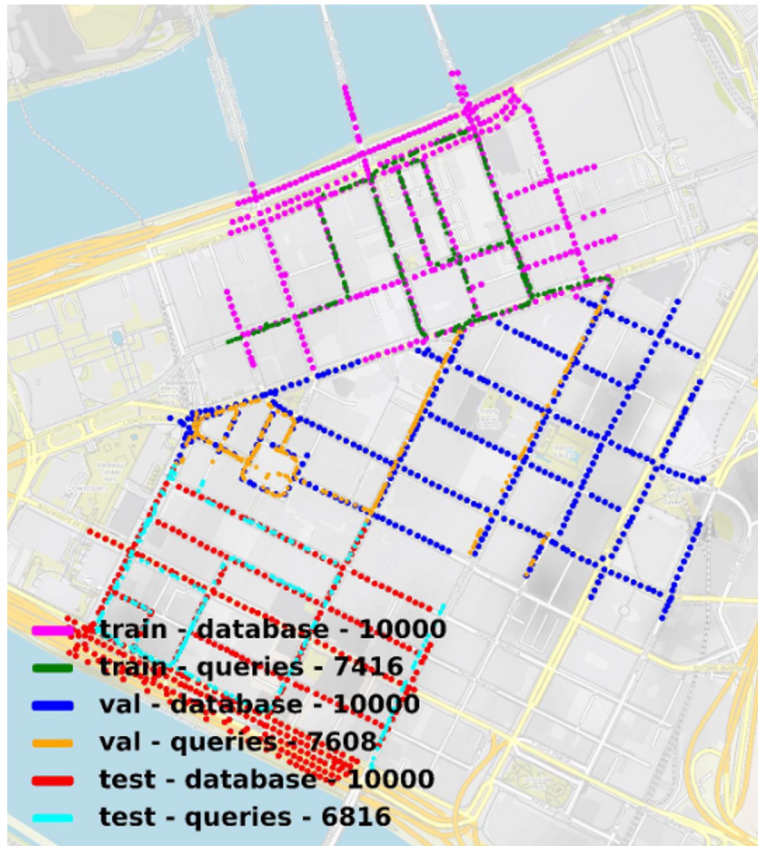
Gabriele Berton
Politecnico di Torino

Towards large- scale applications

Challenges and solutions

Investigating scalability in Geo-localization

Investigation on scalability was limited by the lack of large-scale datasets



Issues:

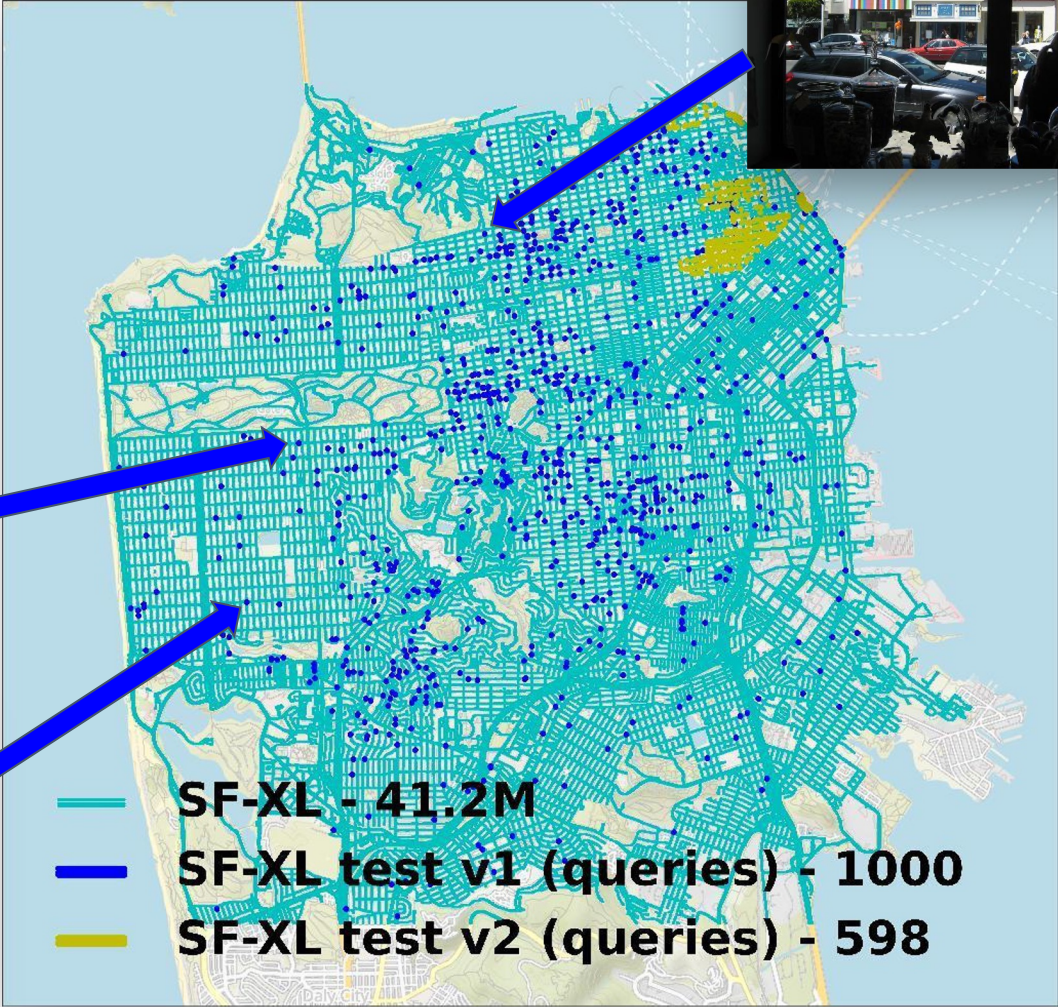
- small scale
- single domain
- non overlapping sets

Solution:

- San Francisco eXtra Large

San Francisco eXtra Large (SF-XL)

- a large scale (41M images) dataset
- single database for training and test
- a challenging set of crowd sourced queries





CosPlace

G. Berton, C. Masone, B. Caputo, "*Rethinking visual geo-localization for large-scale applications*", CVPR 2022

Geo-localization on a large dataset

Previous method did not scale to large datasets:

- at train time - due to mining;
- at test time - due to very high-dimensional descriptors.

Geo-localization on a large dataset

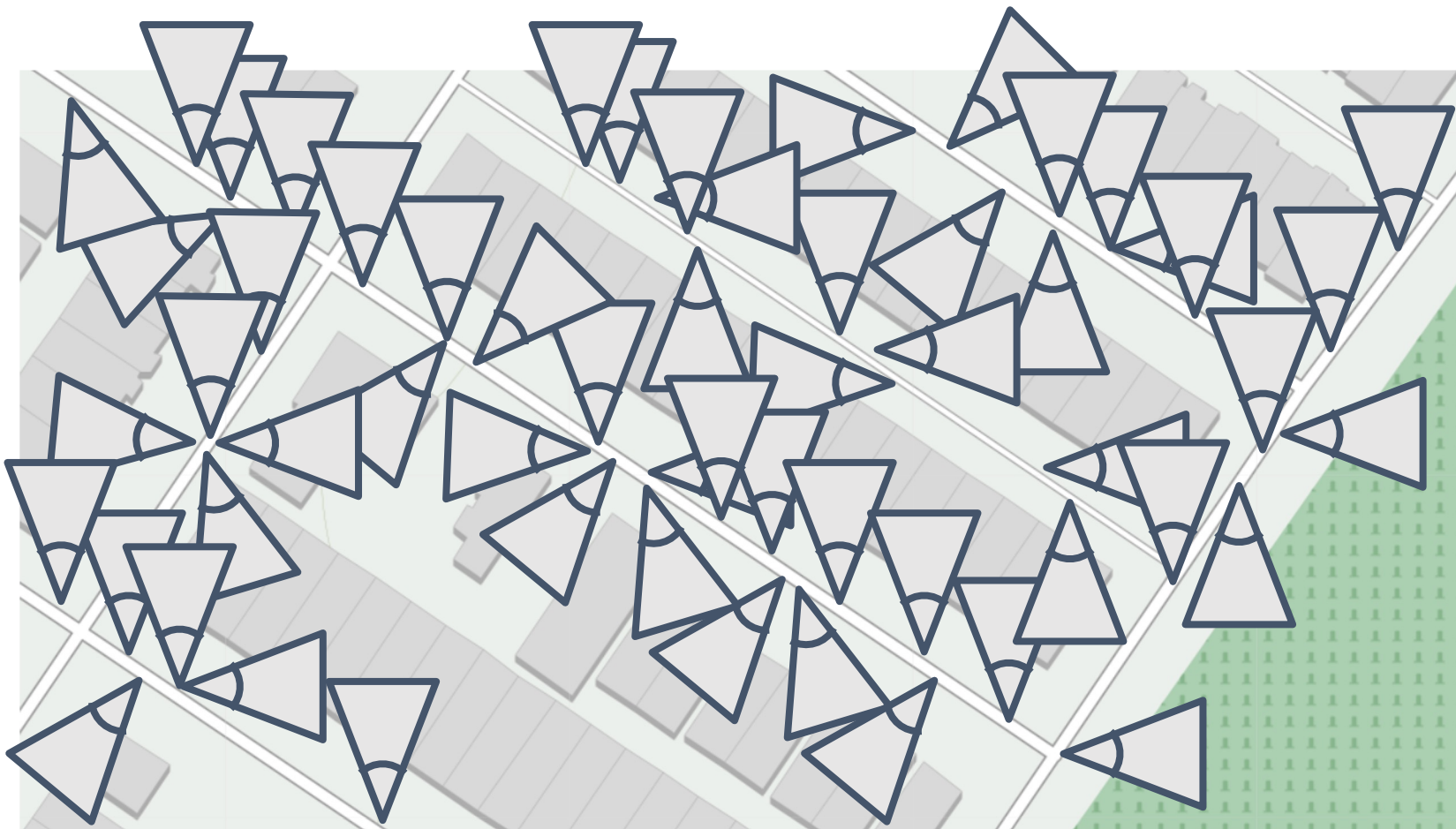
To train on large datasets, we apply best practices from large scale image retrieval (e.g. Google Landmark Retrieval Challenge)

- use scalable losses such as CosFace
- have the images split into classes

Datasets used in Visual Geo-localization have continuous labels (GPS coordinates), and can't be easily split into classes

Our new goal: splitting the dataset into classes

Naive approach for dataset split



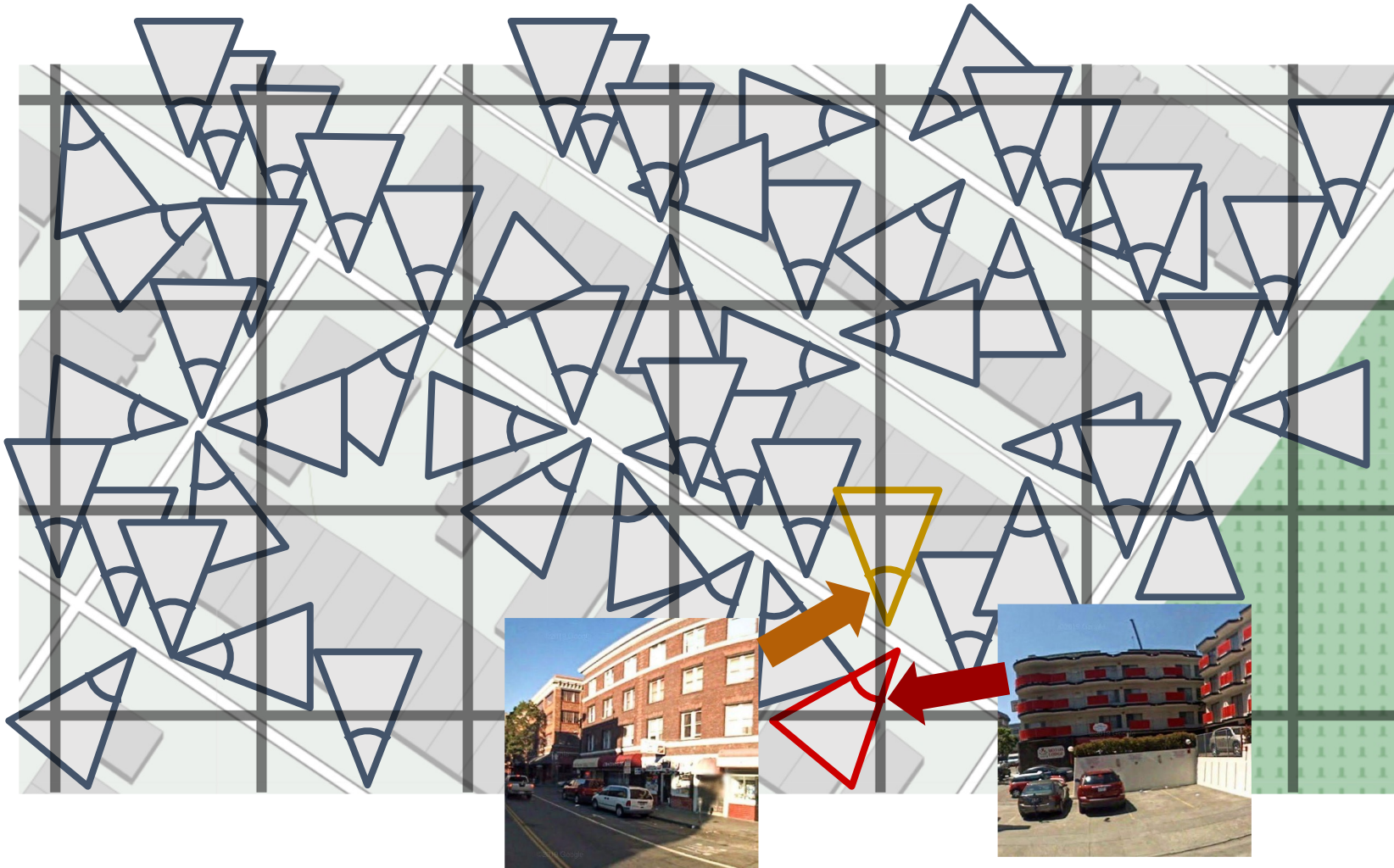
Naive approach for dataset split



Naive approach for dataset split



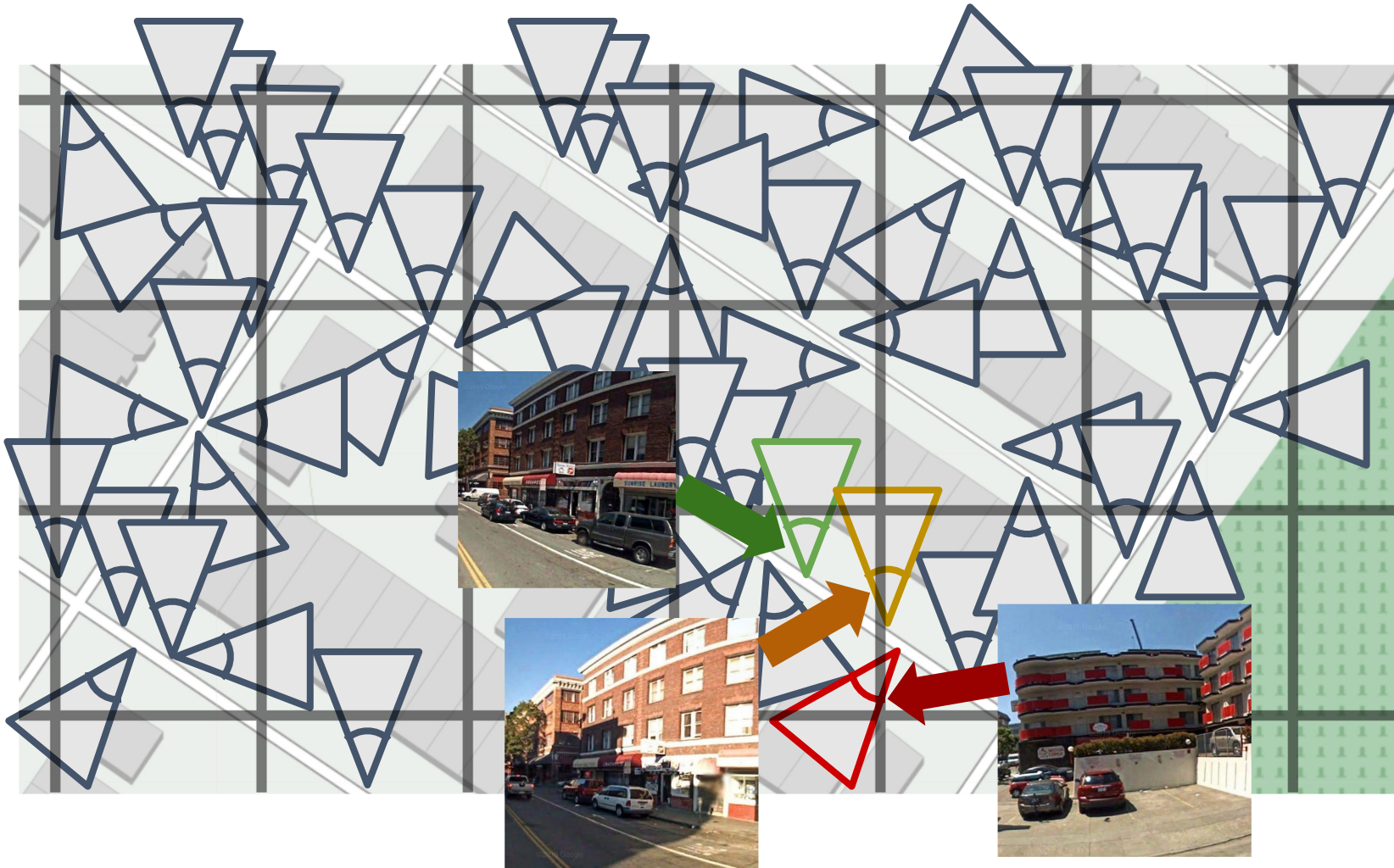
Naive approach for dataset split



Naive split has 2 issues:

1. intra-class scene variation

Naive approach for dataset split



Naive split has 2 issues:

1. intra-class scene variation
2. inter-class visual overlap

Splitting the dataset into classes



Naive split has 2 issues:

1. intra-class scene variation
- Solution: keep only images pointing north
2. inter-class visual overlap

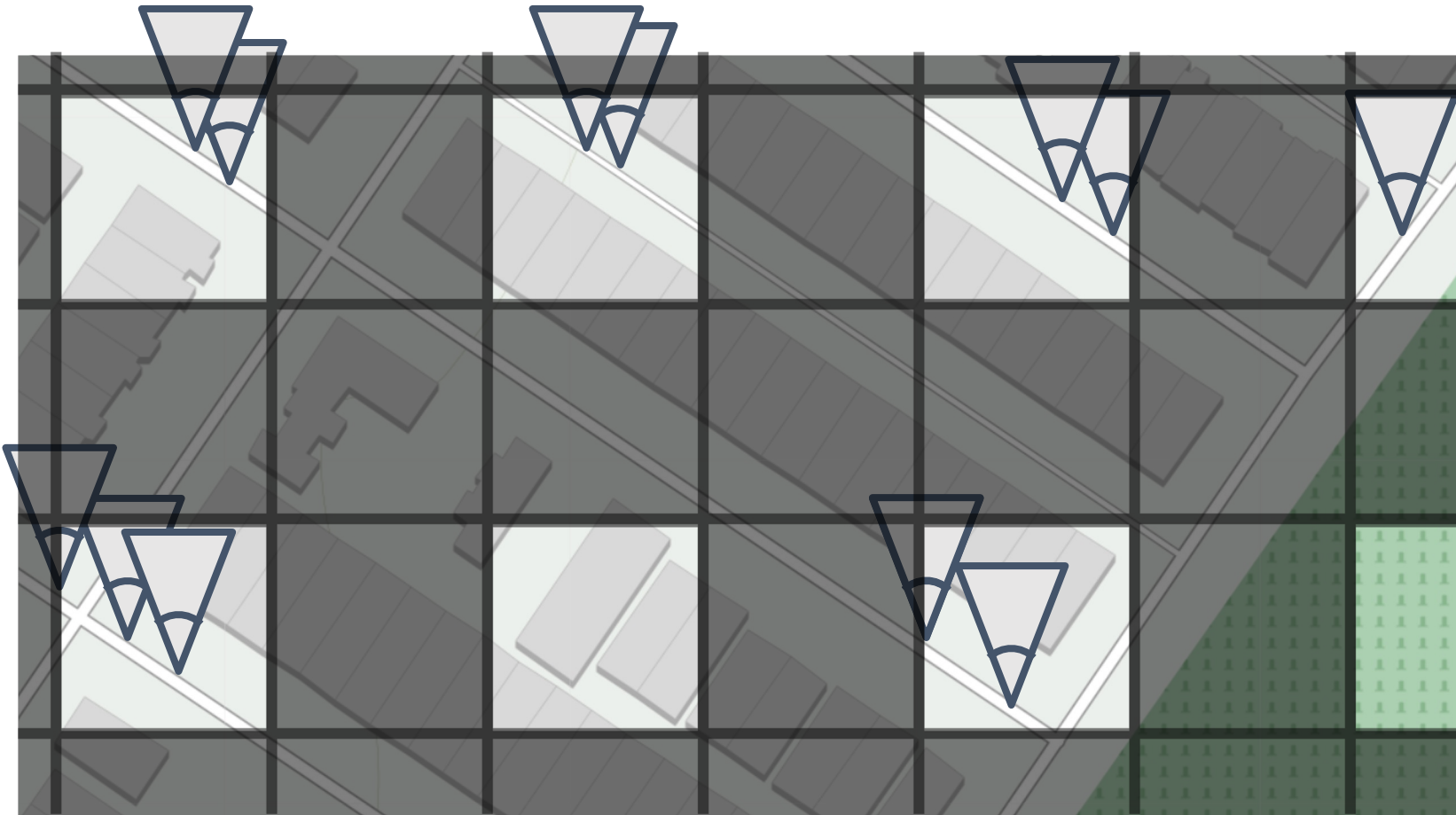
Splitting the dataset into classes



Naive split has 2 issues:

1. intra-class scene variation
 - Solution: keep only images pointing north
2. inter-class visual overlap
 - Solution: remove adjacent classes

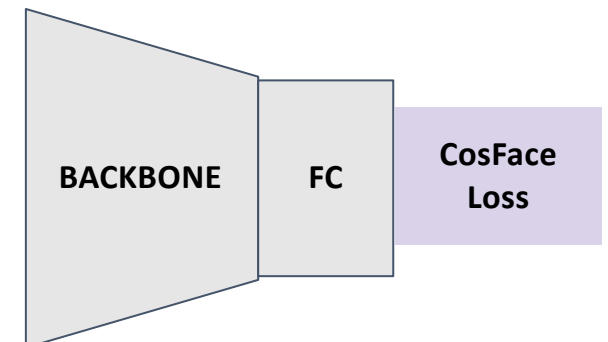
Splitting the dataset into classes



Now the dataset is split into classes

- Train with CosFace
- Obtain SOTA!

Can we go even further?



Splitting the dataset into classes

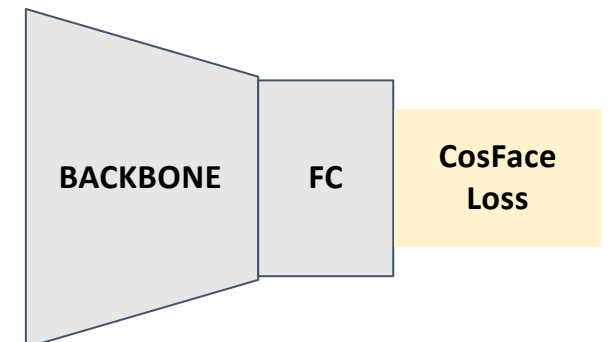


Now the dataset is split into classes

- Train with CosFace
- Obtain SOTA!

Can we go even further?

shift the grid, re-initialize CosFace,
and train more with new classes



Splitting the dataset into classes

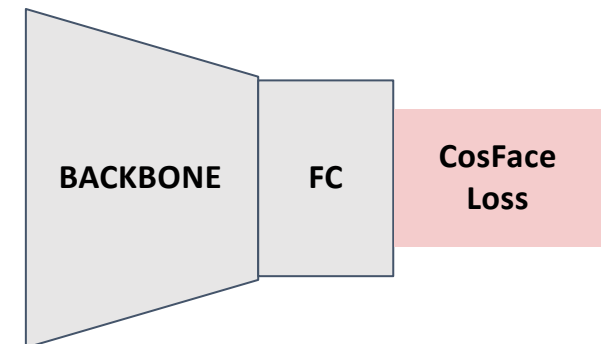


Now the dataset is split into classes

- Train with CosFace
- Obtain SOTA!

Can we go even further?

shift the grid, re-initialize CosFace,
and train more with new classes



Splitting the dataset into classes

CosPlace

Dataset is split into

CosFace

!A!

en further?

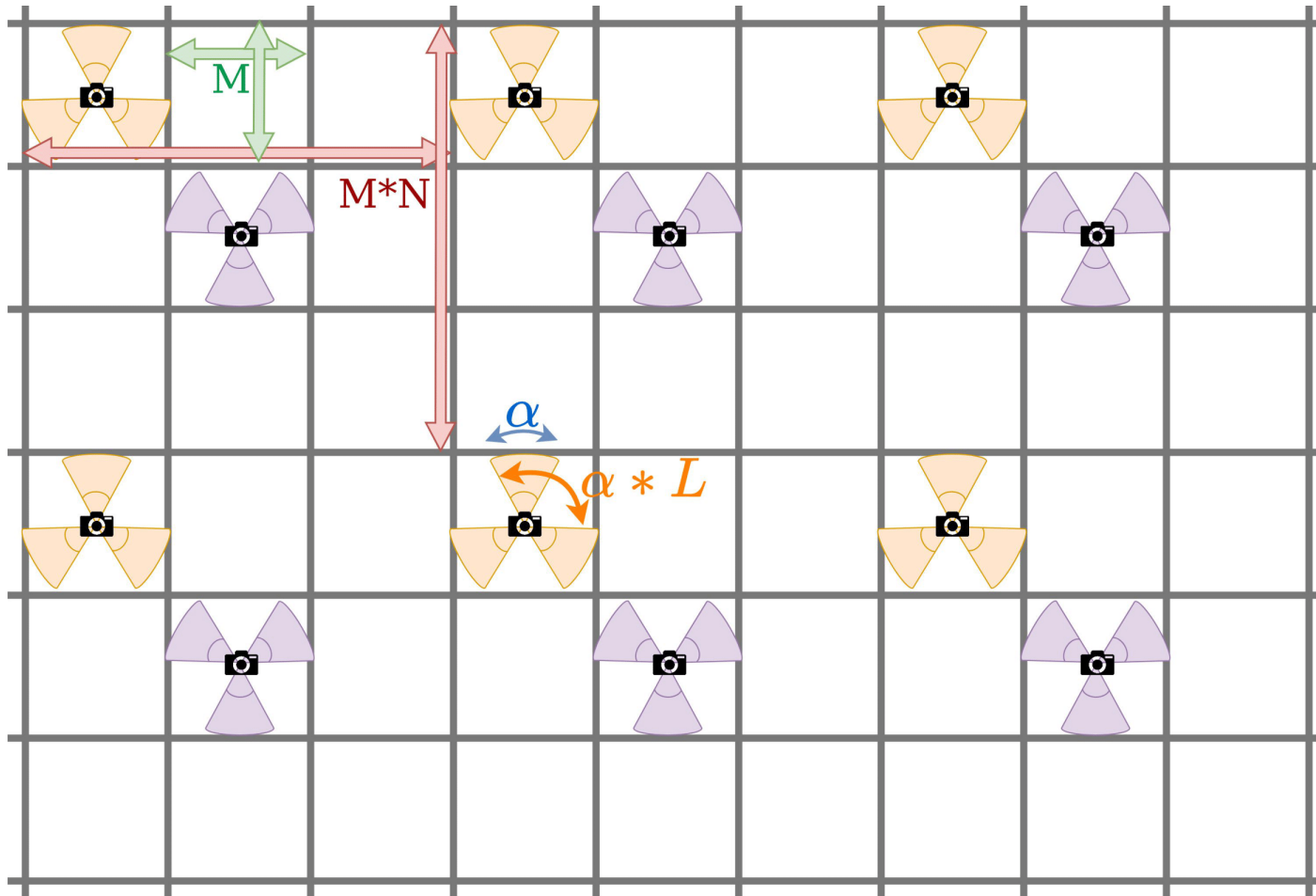
re-initialize CosFace,
e with new classes

BACKBONE

FC

CosFace
Loss

CosPlace



- M – cell side – 10 meters
- N – distance between 2 cells of same subset – $5 \times M$
- α - class FOV - 30°
- L – distance ($^\circ$) between 2 classes of same subset - $2x\alpha$

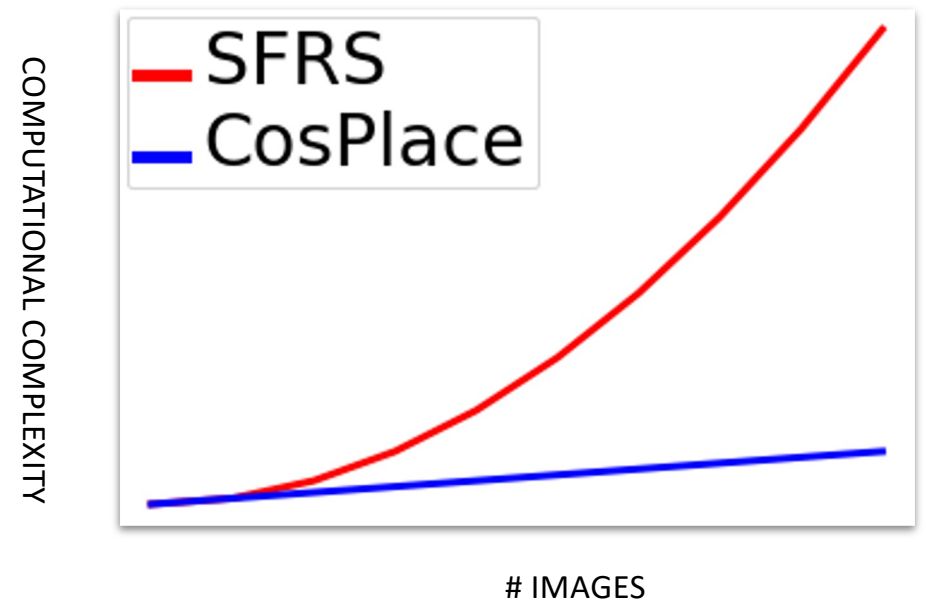
Now some
results



Results 1: previous methods do not scale

Training on San Francisco eXtra Large

Method	Recall @ 1
GeM [1]	21.7
NetVLAD [2]	38.3
SARE [3]	-
SFRS [4]	-
CosPlace [5]	64.7



[1] Radenovic et al, Fine-tuning CNN Image Retrieval with No Human Annotation, PAMI 2018

[2] Arandjelovic et al, NetVLAD: CNN architecture for weakly supervised place recognition, PAMI 2017

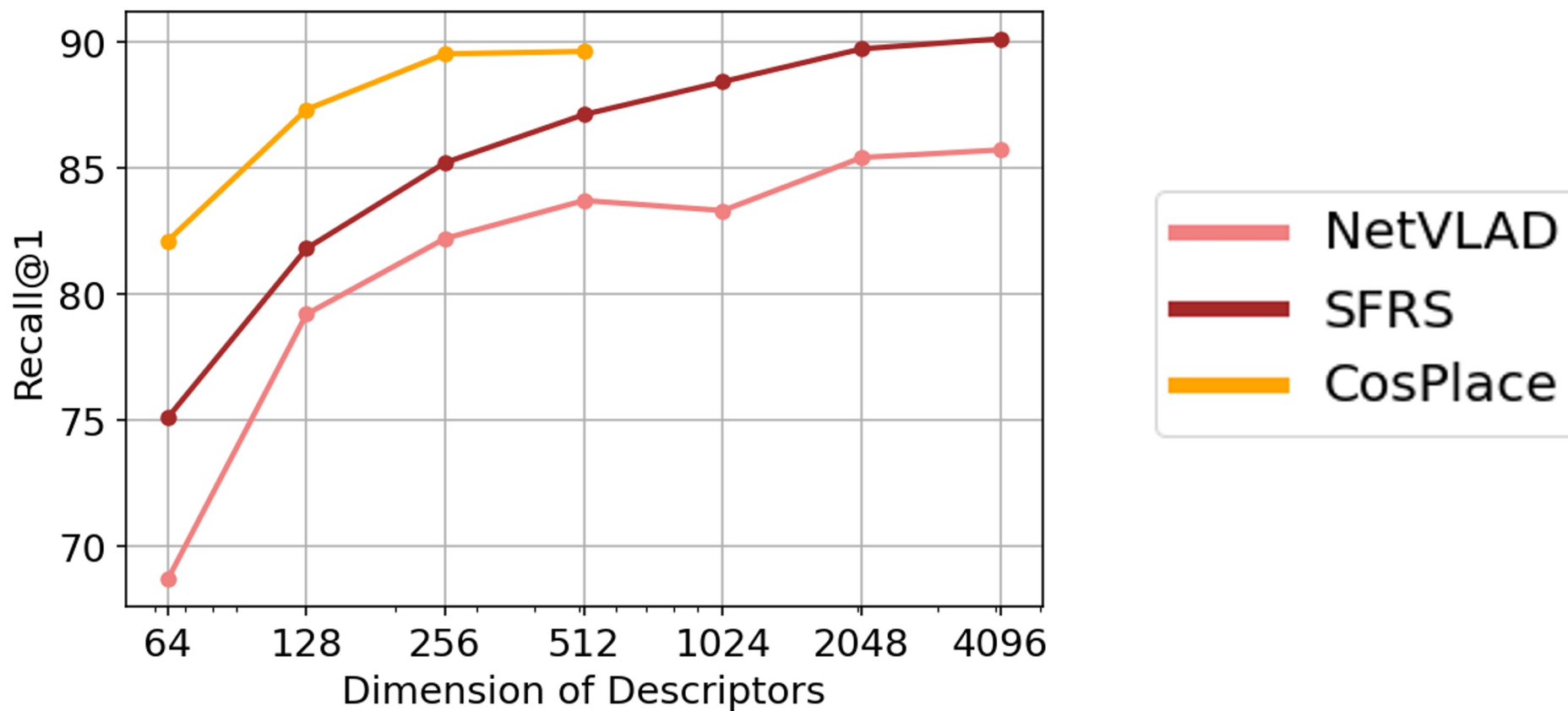
[3] Liu et al, Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization, ECCV 2019

[4] Ge et al, Self-supervising fine-grained region similarities for large-scale image localization, ECCV 2020

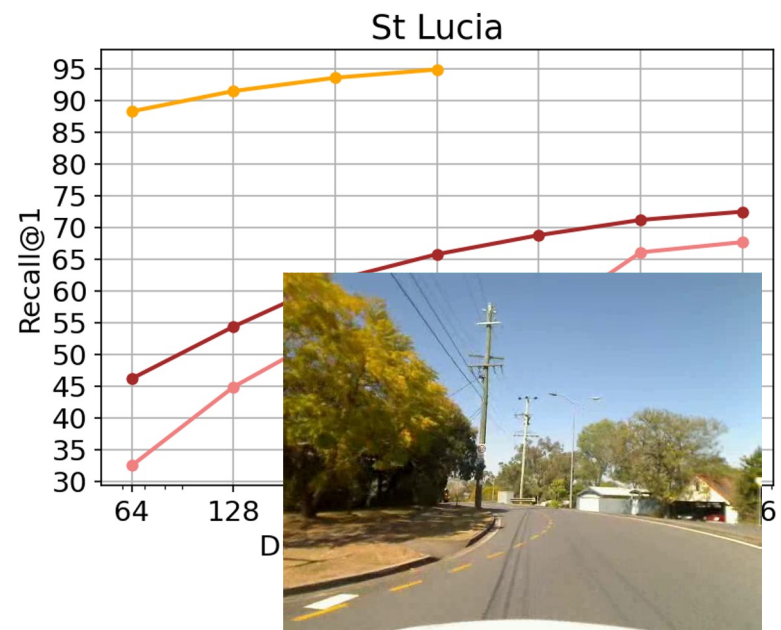
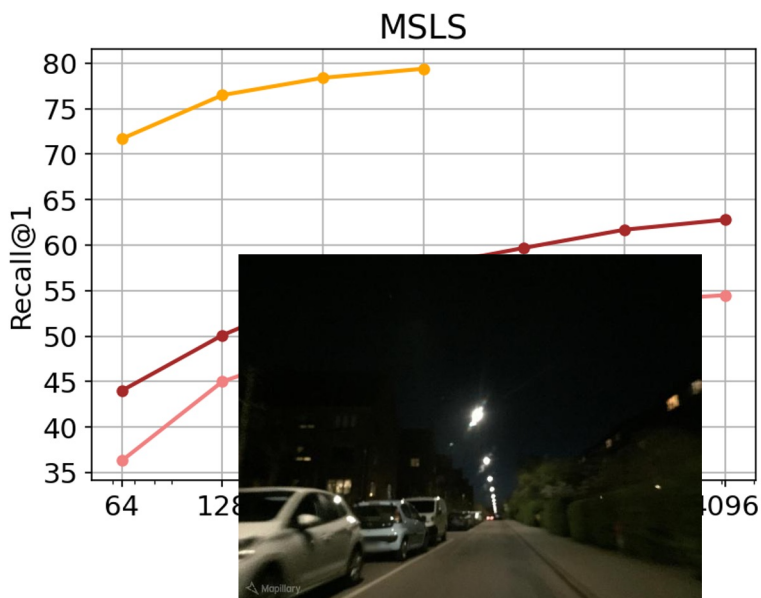
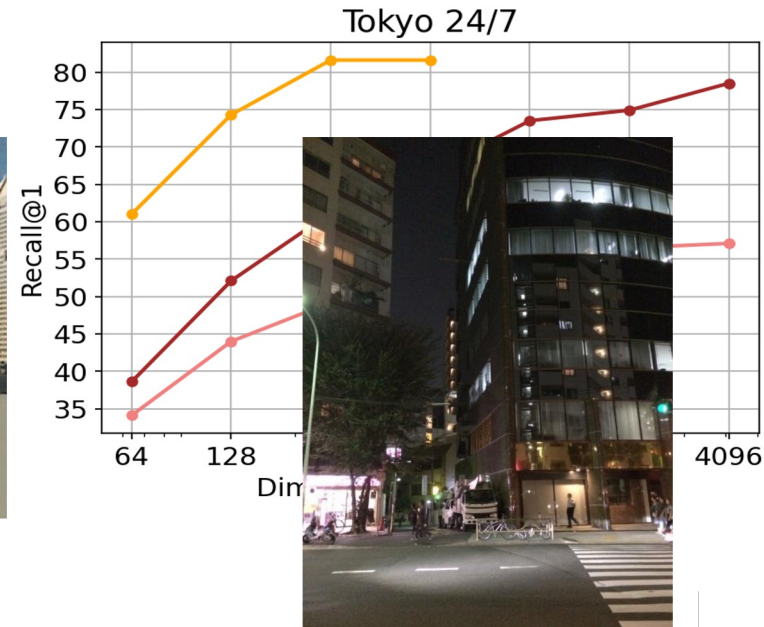
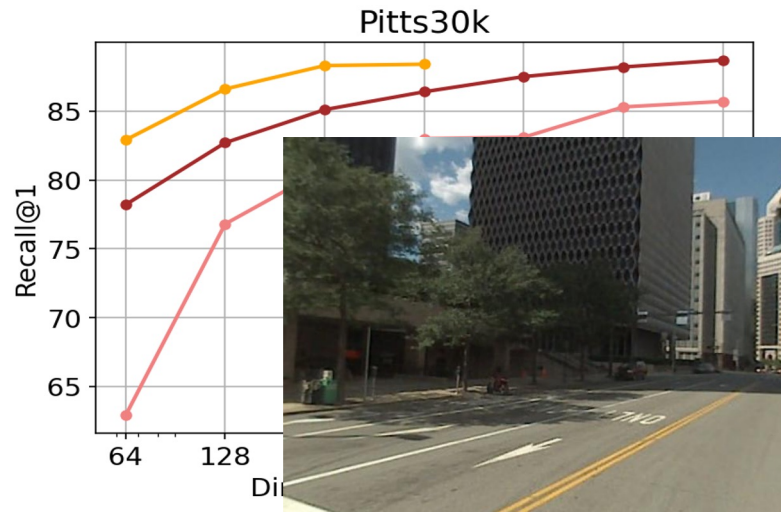
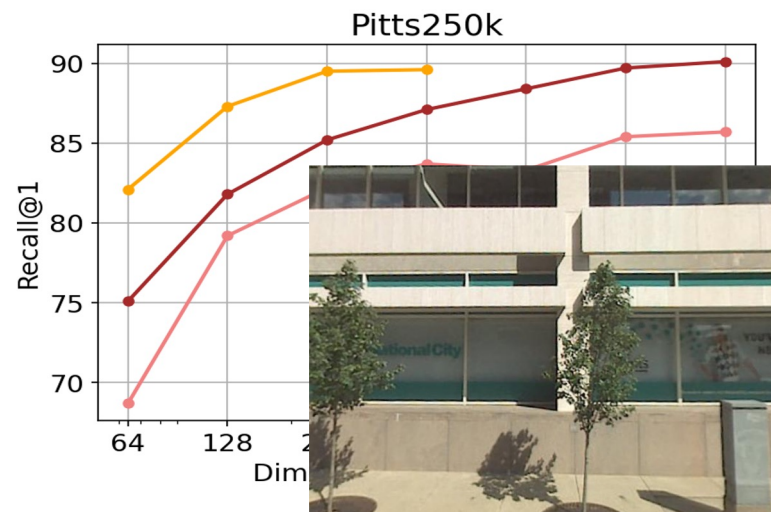
[5] Berton et al, Rethinking Visual Geo-localization for Large-Scale Applications, CVPR 2022

Results 2: CosPlace is lightweight

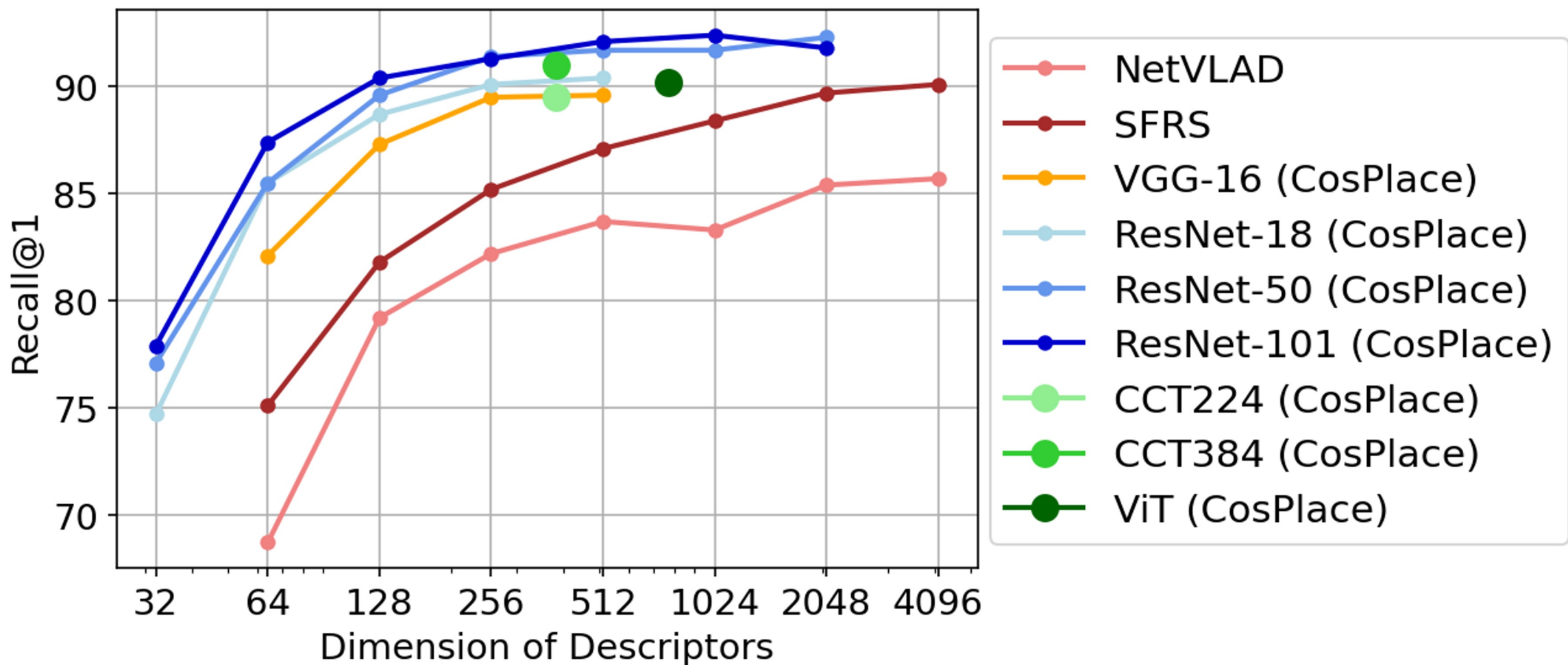
Test on Pittsburgh 250k



Results 3: CosPlace is robust to new cities



Results 4: CosPlace is robust to backbone changes



Advantages and disadvantages of CosPlace

- ✓ outperforms previous SOTA with 8x smaller descriptors
- ✓ training takes only 1 day on 8 GB GPU
- ✓ code is open-source & fully reproducible!
- ✗ requires heading labels
- ✗ unsuitable for training on small datasets

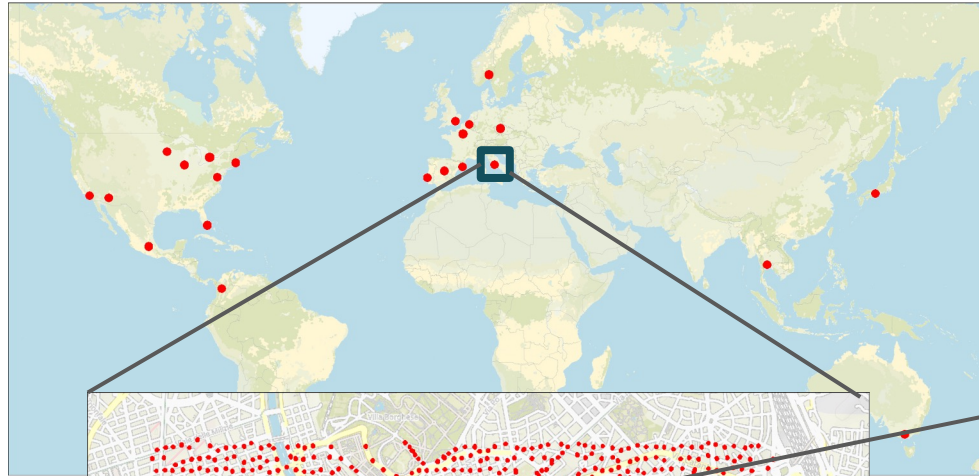
A high-angle, wide shot of a densely packed residential neighborhood in winter. The houses are built on a hillside, with snow covering the roofs and the ground. The houses are painted in various colors, including white, yellow, red, blue, and grey. The roofs are mostly gabled and covered in snow. The sky is overcast and grey. The overall scene is a typical winter landscape in a residential area.

GSV-cities

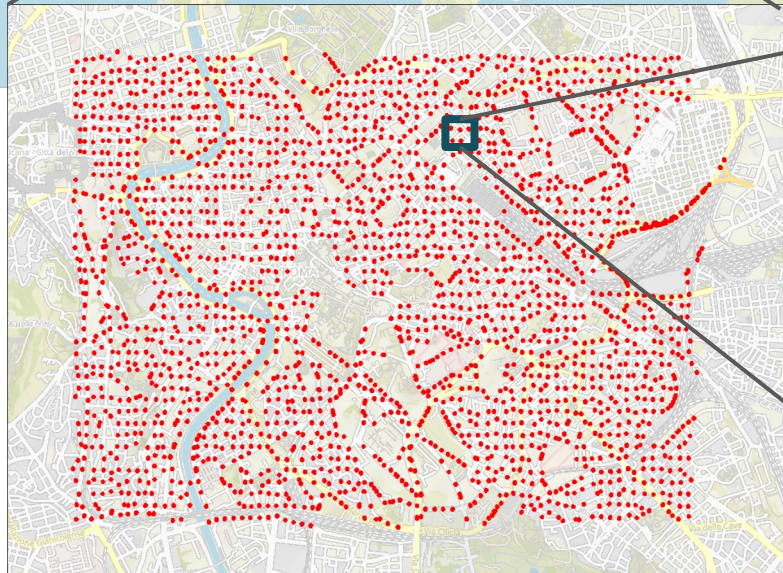
a new dataset

A. Ali-bey et al., "*GSV-Cities: Toward Appropriate Supervised Visual Place Recognition*", Neurocomputing 2022

GSV-cities - a new dataset



- multi-cities and large-scale
- split in classes offline
- only train set



GSV-cities - a new dataset

- mapping geo-localization to standard retrieval
- clear boundary among classes
- no database-queries split
- best results with a multi-similarity loss [1]

GSV-cities - a new dataset

Training with two proposed aggregation layers

Method	Pitts30k	Tokyo 24/7	SF-XL v1	MSLS
NetVLAD [1]	84.6	85.5	40.0	60.1
SFRS [2]	89.0	81.3	50.8	70.0
CosPlace [3]	90.9	87.3	76.4	87.4
Conv-AP [4]	90.6	76.8	49.1	82.6
MixVPR [5]	91.6	86.3	72.3	88.1

[1] Arandjelovic et al, NetVLAD: CNN architecture for weakly supervised place recognition, PAMI 2017

[2] Ge et al, Self-supervising fine-grained region similarities for large-scale image localization, ECCV 2020

[3] Berton et al, Rethinking Visual Geo-localization for Large-Scale Applications, CVPR 2022

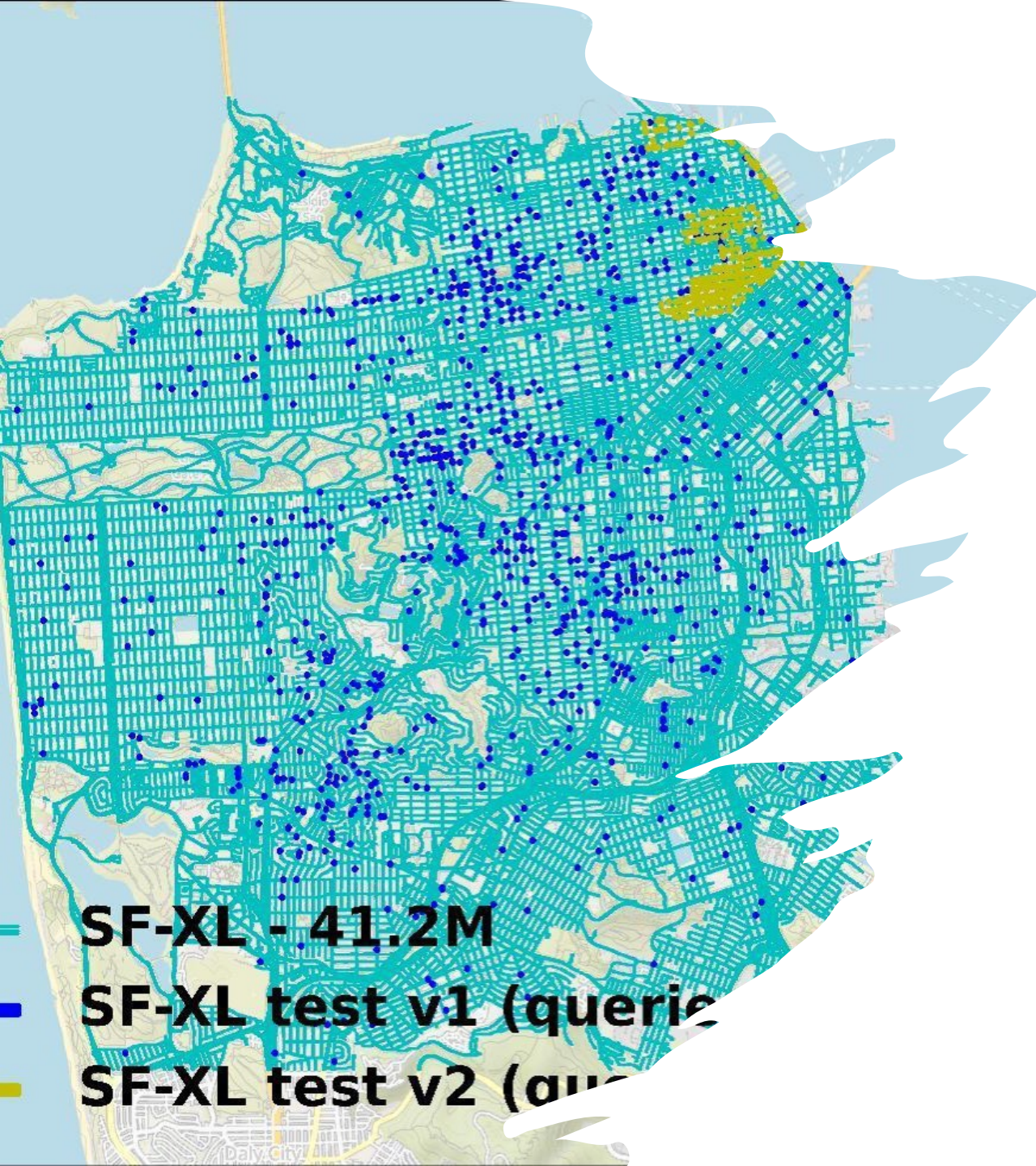
[4] Ali-Bey et al, GSV-Cities: Toward Appropriate Supervised Visual Place Recognition, Neurocomputing 2022

[5] Ali-Bey et al, MixVPR: Feature Mixing for Visual Place Recognition, WACV 2023

Conclusions

What we can achieve
&
open challenges





SF-XL - 41.2M

SF-XL test v1 (queries)

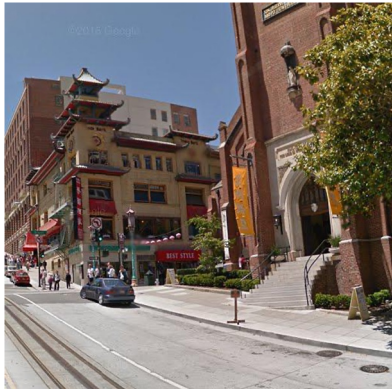
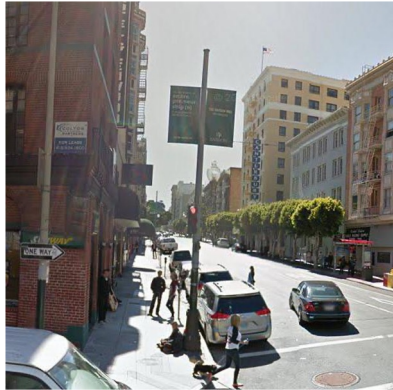
SF-XL test v2 (queries)

Geo-localization capabilities

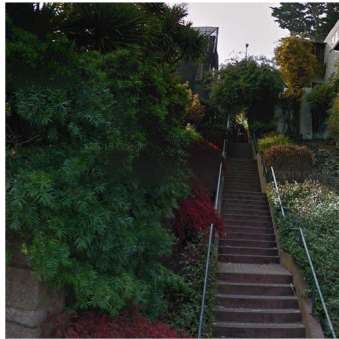
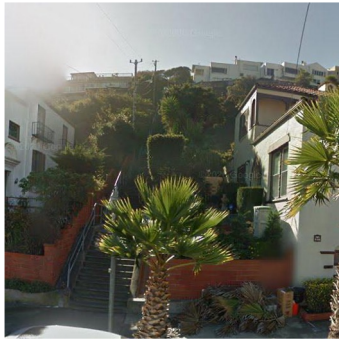
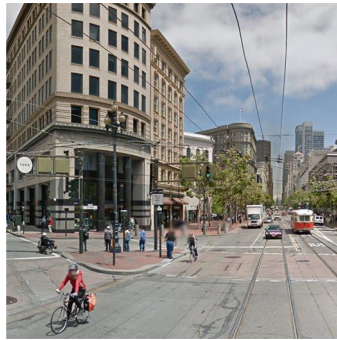
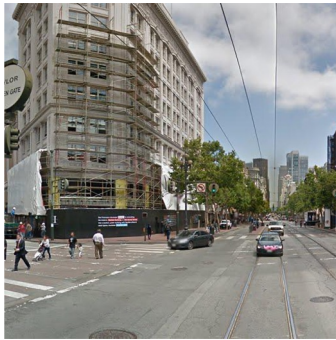
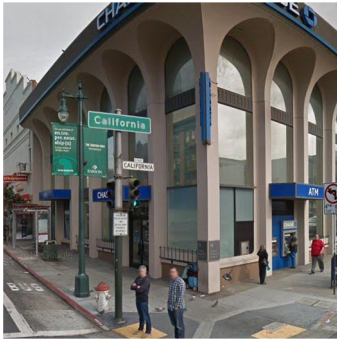
Qualitative results on San Francisco XL

- 2.8 M images of database
- ~ 150 sq km
- 512-D descriptors
- ~ 8GB RAM
- 500 ms for 1 query
 - 25 ms/q with batched kNN
 - 1 ms/q with approximate kNN

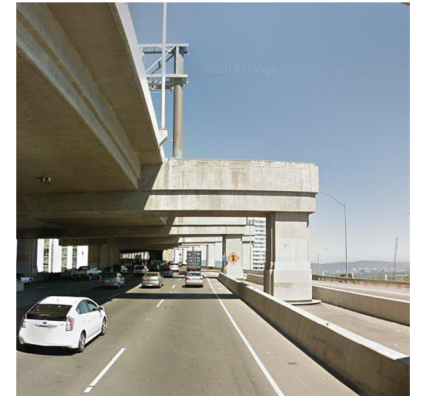
Correct predictions



Wrong predictions - Query, Positive, Prediction



Wrong predictions - outliers





Open question: scalability

- how far can we scale with image retrieval?
- can compact descriptors discriminate between millions of places?
- hard to tell due to lack of bigger datasets

My guess on the future

- Scaling up will require multi-step solutions
 - classification + retrieval + re-ranking
- We won't see bigger datasets for a few years
- CNNs will still outperform transformers
- Domain adaptation will be helpful



Geolocalize
movies:
Venom



37.79789 -122.41371



Geolocalize
movies:
Shang Chi



37.78961 -122.41291



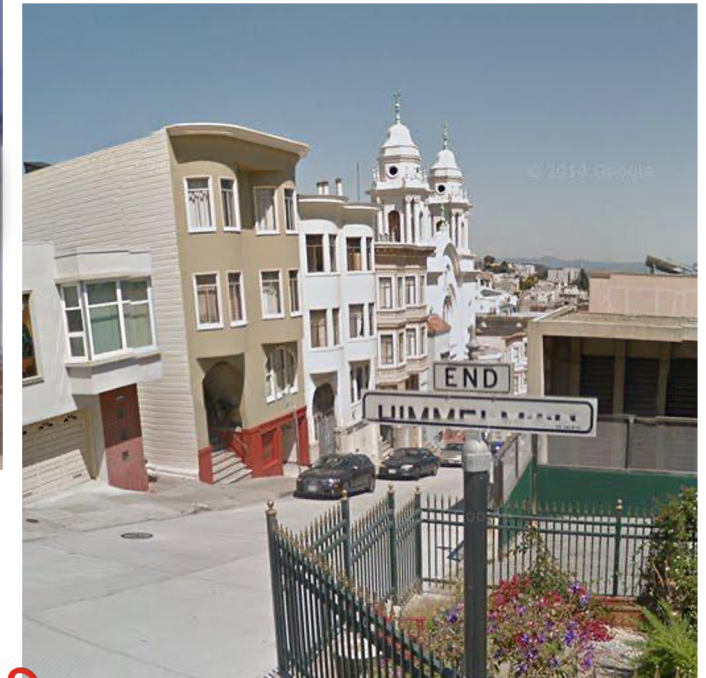
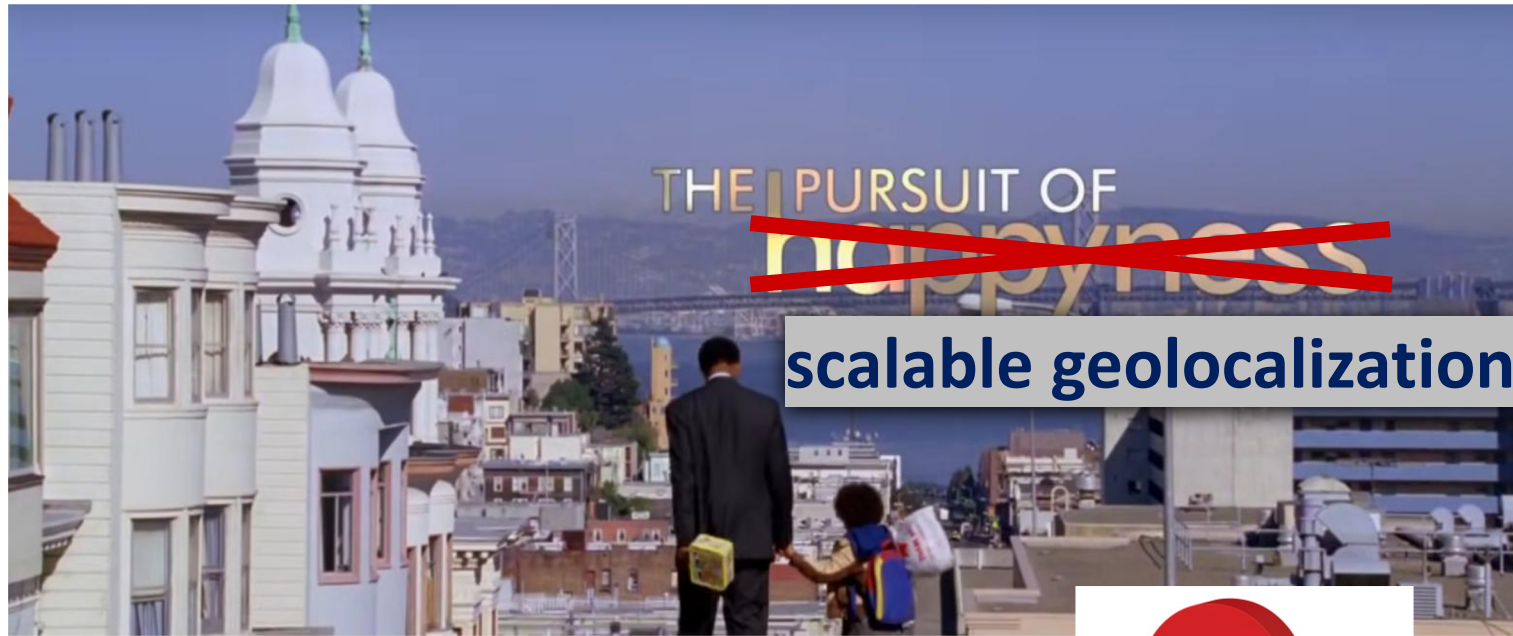
3D Demo

The image displays a 3D demo interface. On the left, a 'Select Image' window is open, showing a file explorer view of a folder named 'Berlin Web'. The window contains a grid of 20 image thumbnails, each with a filename. The files are:

- 2006_08_07_Rathaus_Schoeneberg.jpeg
- 8873_d7ced_577cd7867120a.jpeg
- 8874_ba53c_577cd72705968.jpeg
- 32155_6b0bf_588e1cb03b653.jpg
- 32155_f9f0a_588e1d87753f6.jpeg
- 366567_746f4_5926dc25a85ba.jpg
- 367662_166ff_592958f9b3c2.jpg
- 401901_c62ee_592daa757e553.jpg
- 402260_53b37_592dab5b94e42.jpg
- 420457_6aca3_592f2361e1d7c.jpg
- 439486_a49d_593029a6ba735.jpg
- 439508_98ac2_593031ac14d21.jpg
- 464372_b0a65_593829e38f5ba.jpg
- 467052_af2eb_593ea3ad80dd0.jpg
- 467053_28871_593ea4e048a03.jpg

At the bottom of the file explorer, there is a 'File name:' field and 'Open' and 'Cancel' buttons. A blue 'Done' button is visible at the bottom left of the 3D view. The 3D view itself shows a cityscape with a prominent building on the right and a 'Query Image' button at the bottom right. The Cesium logo and text 'CESIUM ion · Upgrade for commercial use · Data attribution' are visible at the bottom left.

Thank you for your attention



Any Questions



37.79707 -122.41300



<https://github.com/gmberton>