Visual similarity learning for instance-level image and video retrieval

Giorgos Kordopatis-Zilos

Retrieval paradigm

query video

video database

what images/videos are considered related?

retrieved videos

$$s(\ \blacksquare\ ,\ \blacksquare\ ) = simillarity$$

how to measure similarity between two images/videos?

Instance-level image retrieval (ILR)

Task: retrieval all images that depicts **a particular object** given a query image or bounding box

query positives

query positives

ROP+1M and GLDv2 datasets

- specific domain
- focused on landmarks

Radenović et al. Revisiting oxford and paris. CVPR, 2018.
Weyand et al. Google landmarks dataset v2. CVPR, 2020.

INSTRE

- multiple domains
- landmarks, planar objects, 3D objects

Wang & Jiang. INSTRE. TOMM, 2015.

Fine-grained Incident Video Retrieval (FIVR)

Task

retrieval all videos of a **particular incident** given a query video

FIVR-200K

- large number of crawled incidents
- 3 video associations
- 3 different tasks

Kordopatis-Zilos et al. “FIVR: Fine-grained Incident Video Retrieval”. IEEE TMM, 2019.

Evaluation metrics

Average Precision (AP)

$$AP = \sum_{i=1}^{N} P(i)(R(i) - R(i-1)) \in [0, 1]$$

mean Average Precision (mAP)

- standard **retrieval** metric
- compute **AP per query**

micro Average Precision (μAP)

- **detection** counterpart
- compute AP with **all query-positive** pairs

Computational requirements

- retrieval time (seconds)
- storage space (MB)

global similarity estimation on images and videos

image: global descriptor extraction

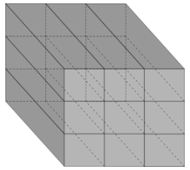
input image

$i \in \mathbb{R}^{H \times W \times 3}$

$f_\theta(\cdot)$

feature map

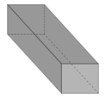
$X \in \mathbb{R}^{h \times w \times D}$



$g(\cdot)$

global descriptor

$\mathbf{x} \in \mathbb{R}^D$

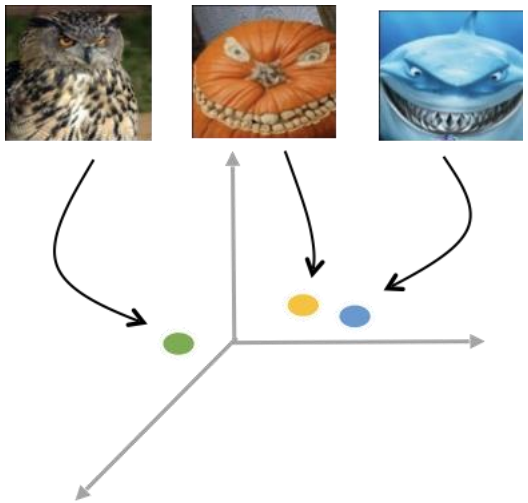


Deep network

- Convolutional Neural Network
- Vision Transformer

global pooling

- average pooling $\mathbf{x} = \sum_{x \in X} x$
- max pooling $\mathbf{x} = \max_{x \in X} x$
- generalized mean pooling $\mathbf{x} = \left(\frac{1}{|X|} \sum_{x \in X} x^p \right)^{\frac{1}{p}}$



Radenovic et al. “Fine-tuning CNN Image Retrieval with No Human Annotation”. TPAMI, 2019.

image: global similarity estimation

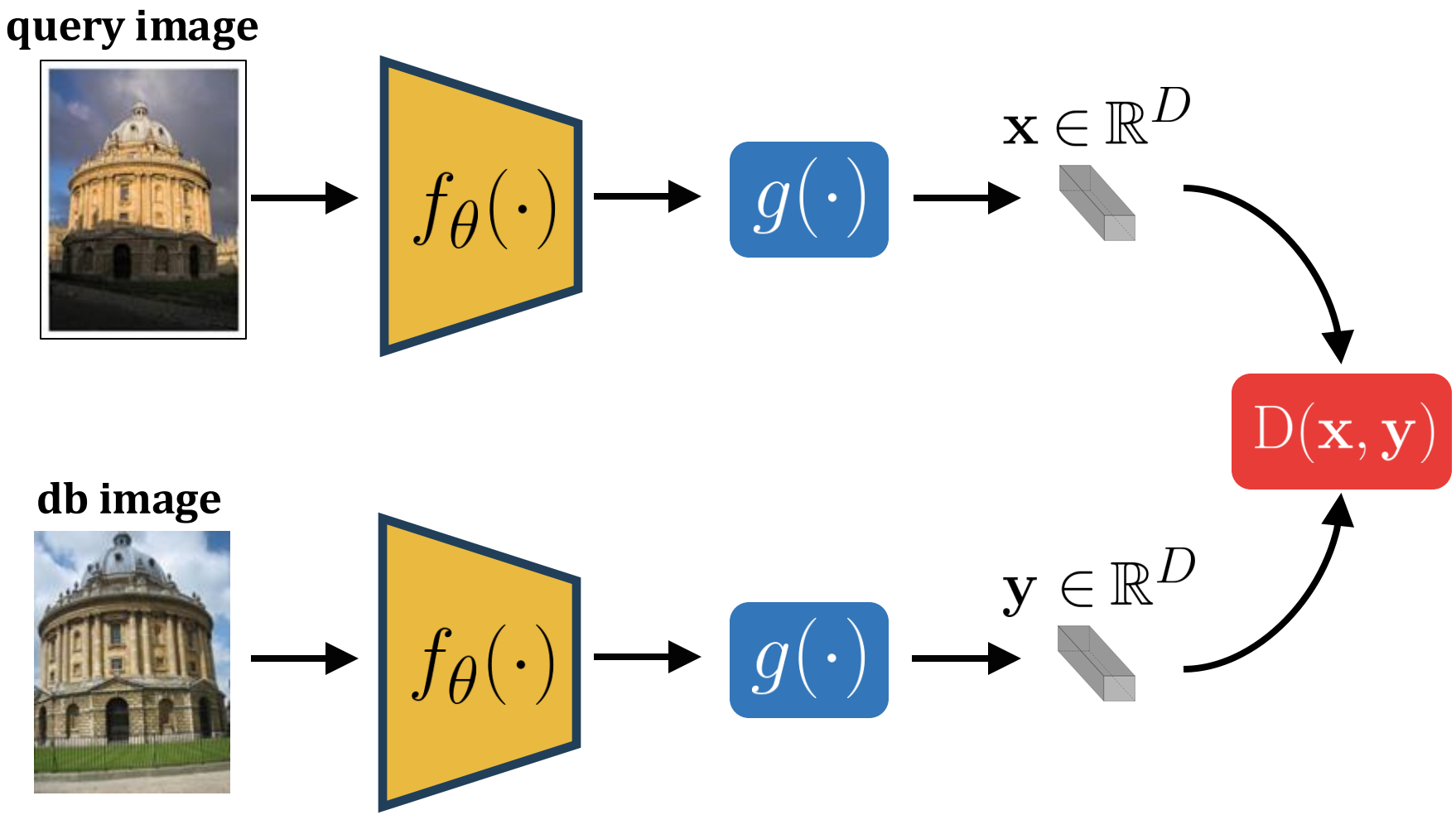
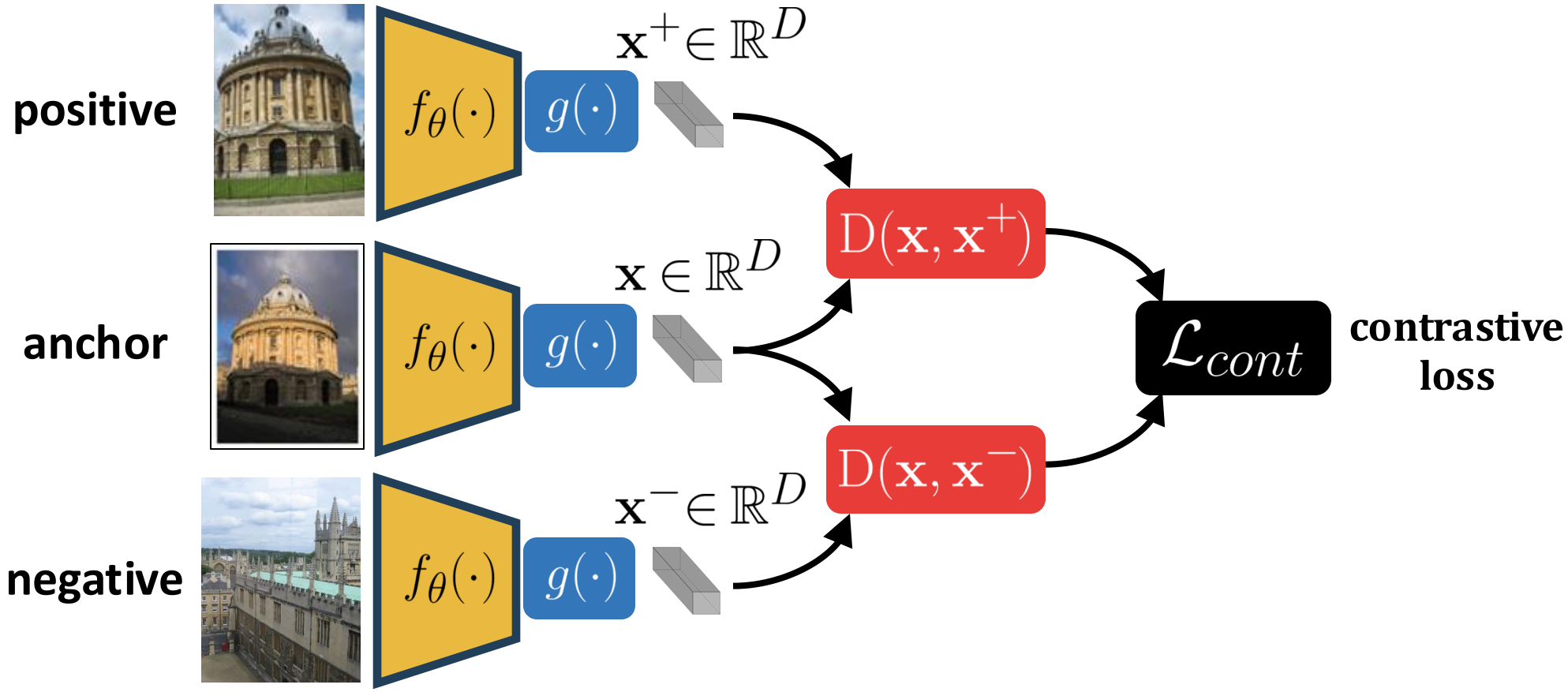


image: global similarity training



Radenović et al. “CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples”. ECCV, 2016.
Lee et al. “Revisiting self-similarity: Structural embedding for image retrieval”. CVPR, 2023.

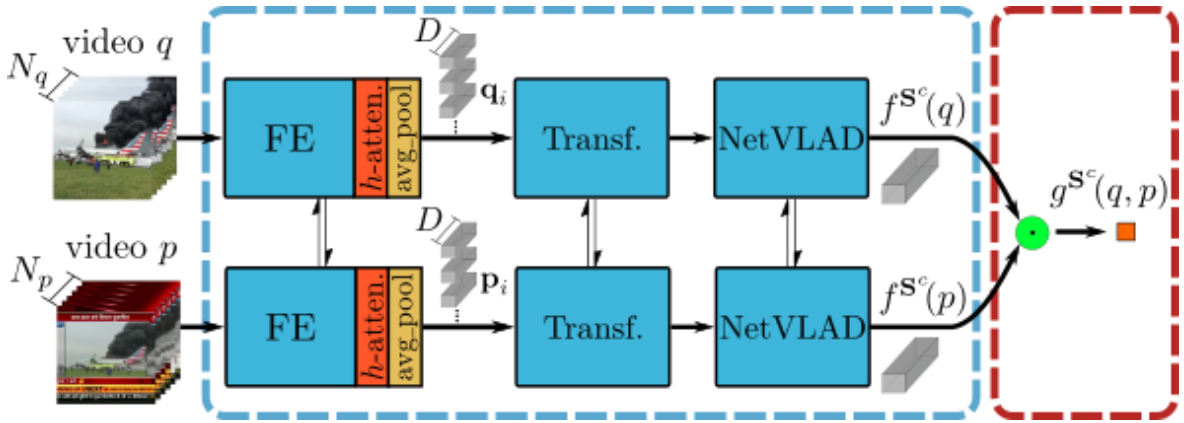
video: global similarity

Feature extraction

- **similar** to image extraction process
- **per frame feature extraction**
  + temporal aggregation

Similarity learning

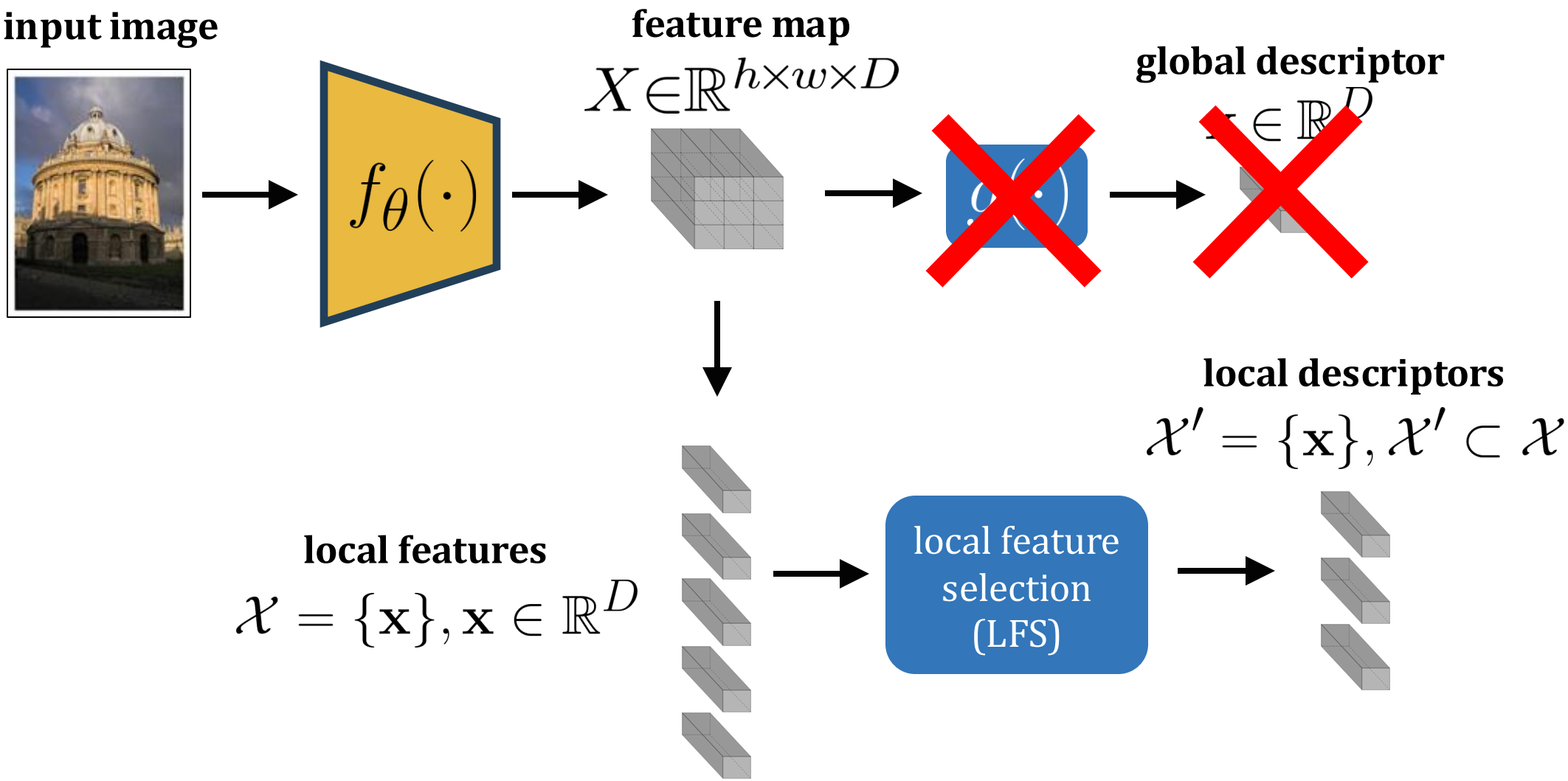
- **similar** to image similarity training
- backbone network remains **frozen**



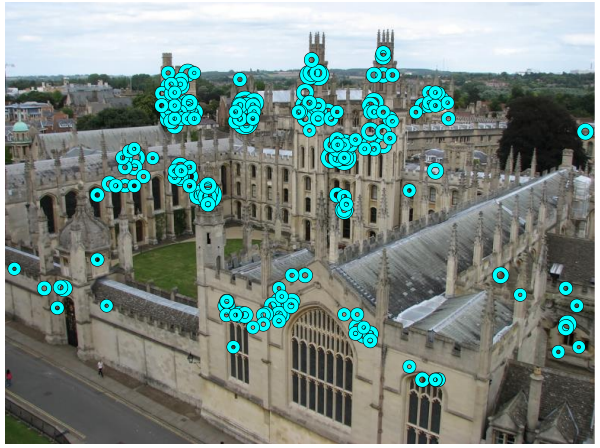
Kordopatis-Zilos et al. "DnS: Distill-and-Select for Efficient and Accurate Video Indexing and Retrieval." IJCV, 2022.
Jo et al. “VVS : Video-to-Video Retrieval with irrelevant Frame Suppression”. AAAI, 2024.

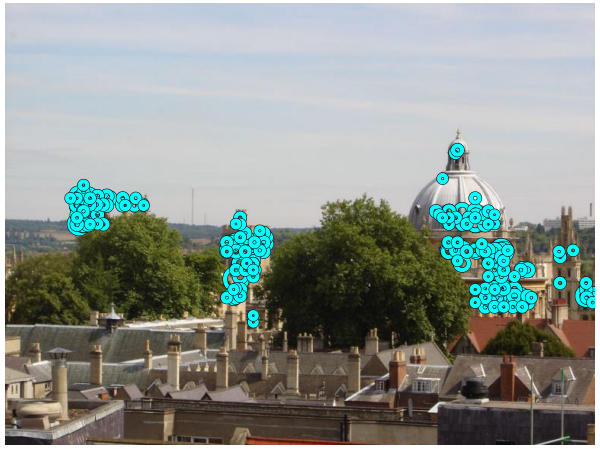
local similarity estimation on images

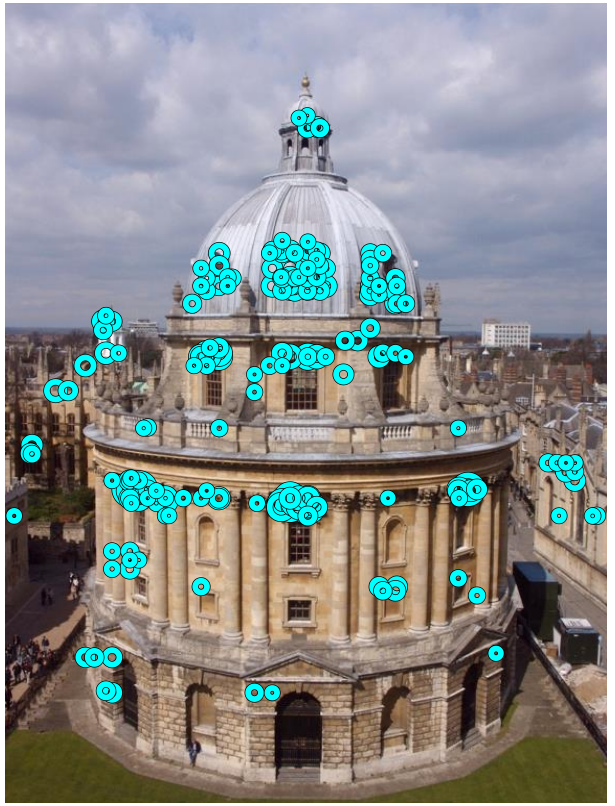
Local descriptor extraction



Local descriptor extraction

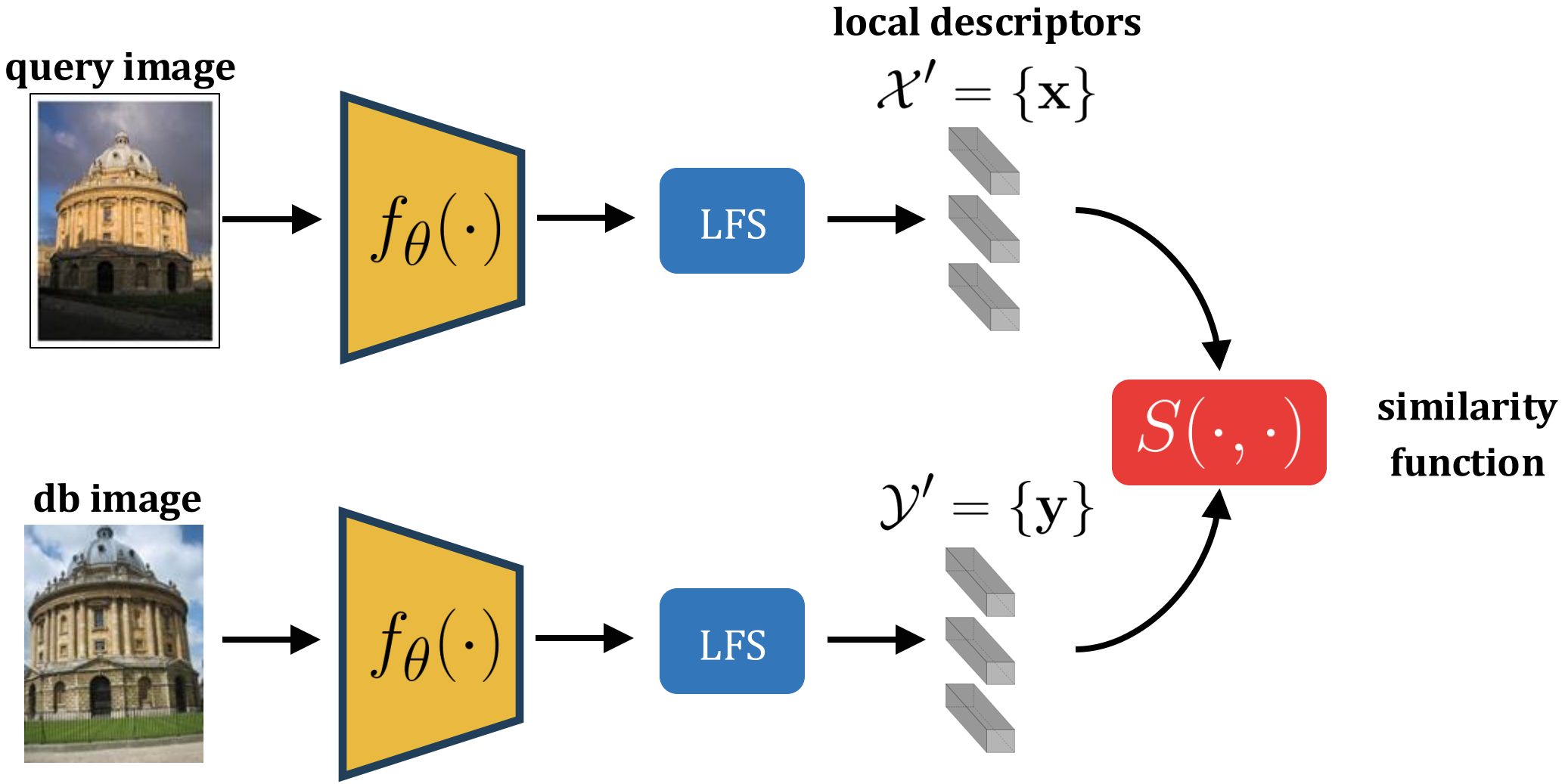






Cao et al. “Unifying deep local and global features for image search”. ECCV, 2020
Tolias et al. “Learning and aggregating deep local descriptors for instance-level recognition”. ECCV, 2020.

Local similarity estimation

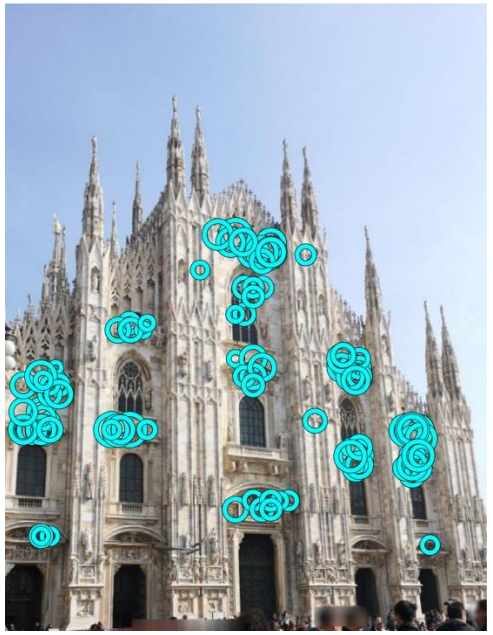


Global similarity	vs	Local similarity
+ very fast		+ high performance
+ memory efficient		- slow
- low performance		- memory inefficient

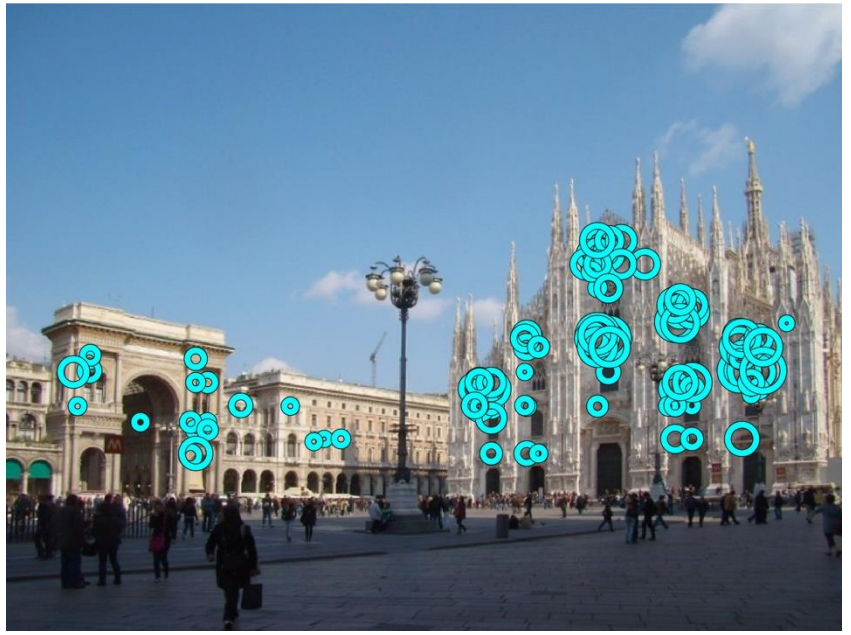
Tan et al. “Instance-level image retrieval using reranking transformers”. ICCV, 2021.
Lee et al. “Correlation verification for image retrieval”. CVPR, 2022.
Zhu et al. “R2former: Unified retrieval and reranking transformer for place recognition”. CVPR, 2023.

Asymmetric similarity

query image
100 descriptors

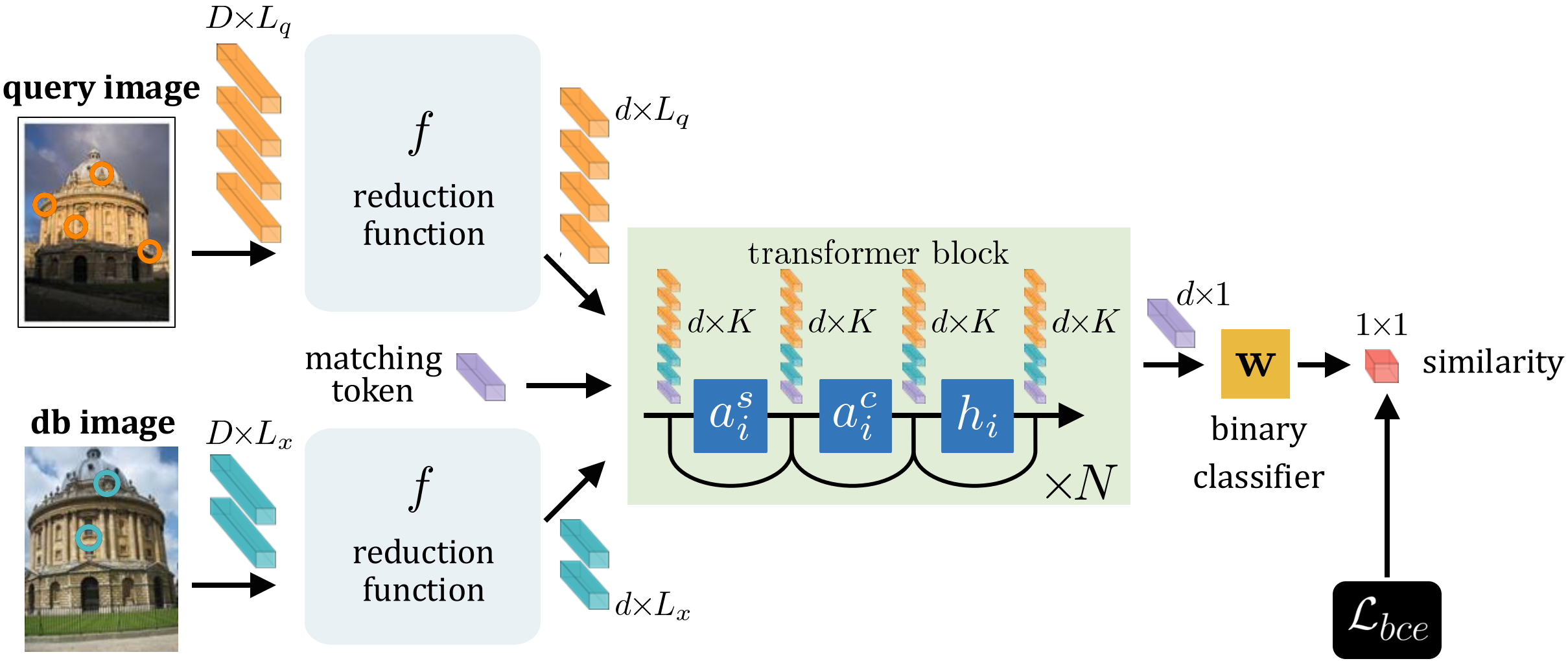


db image
100 descriptors



Suma et al. “AMES: Asymmetric and Memory-Efficient Similarity Estimation for Instance-level Retrieval”. ECCV, 2024.

AMES: Asymmetric Memory-Efficient Similarity

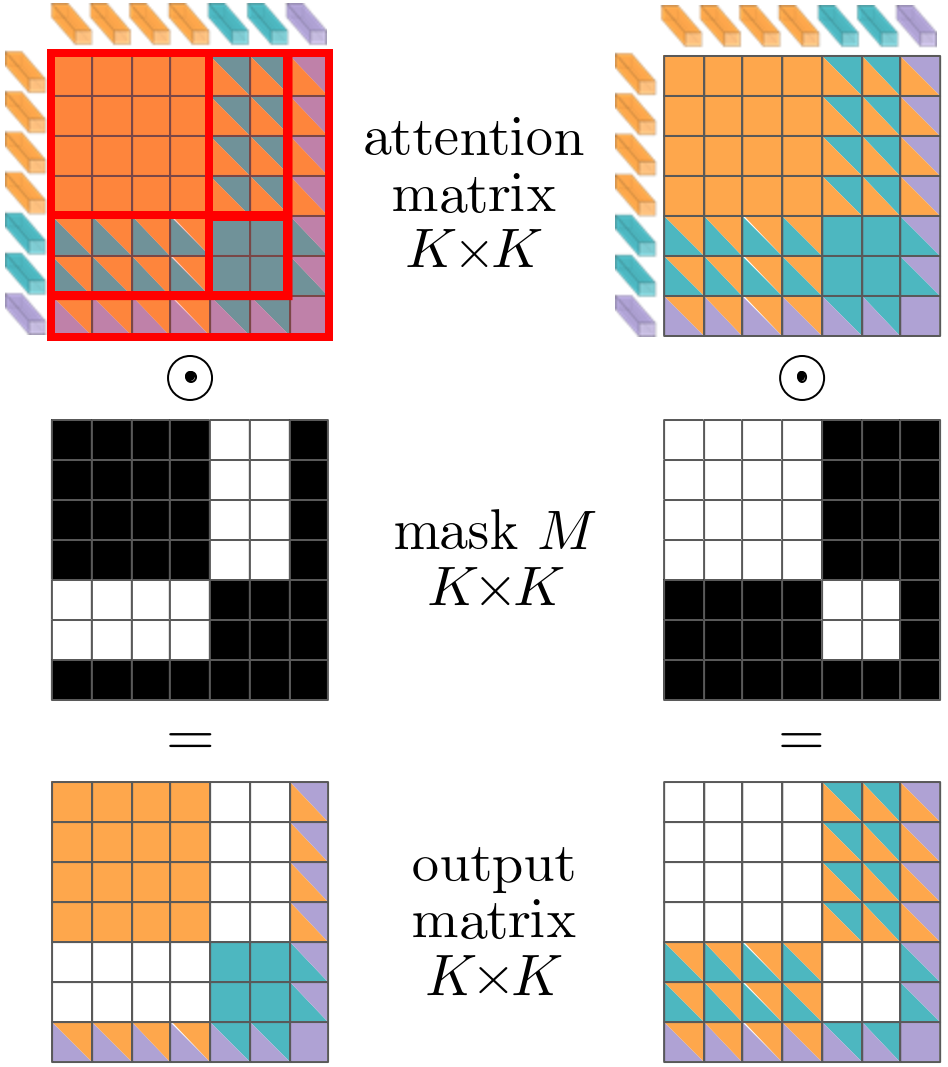


Suma et al. "AMES: Asymmetric and Memory-Efficient Similarity Estimation for Instance-level Retrieval". ECCV, 2024.

AMES attention layers

a_i^s - **self-attention**

- **intra-image** attention
- tokens attends only to others from **the same image** and matching token



a_i^c - **cross-attention**

- **inter-image** attention
- tokens attends only to others from **the other image** and matching token

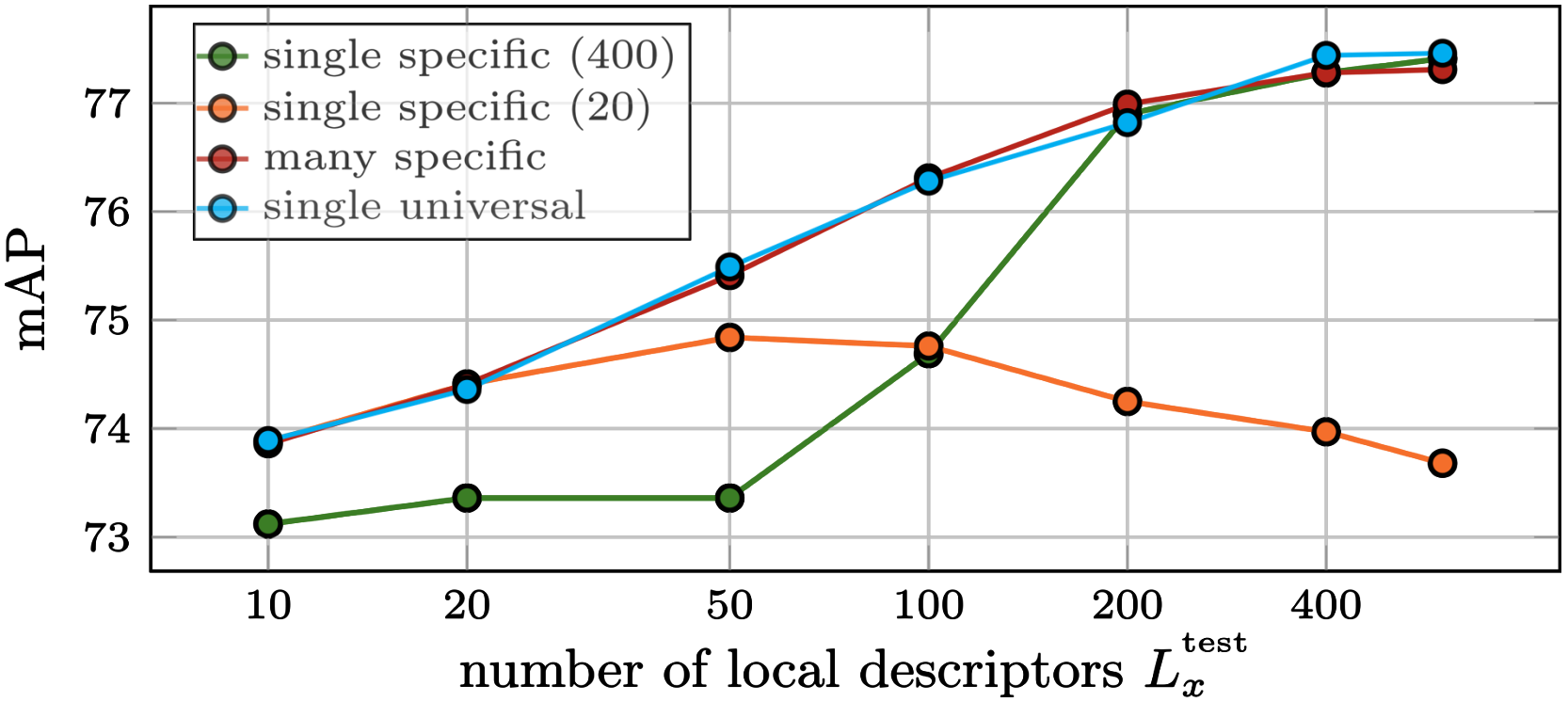
Universal training

Problem

- transformers are senitive to discrepancy in input sequence length

Solution

- sample a random number local descriptors per image during training



Global-local similarity ensemble

query image

initial ranking w/ global similarity $s_g(x, q)$

top-K nearest neighbours

re-rank

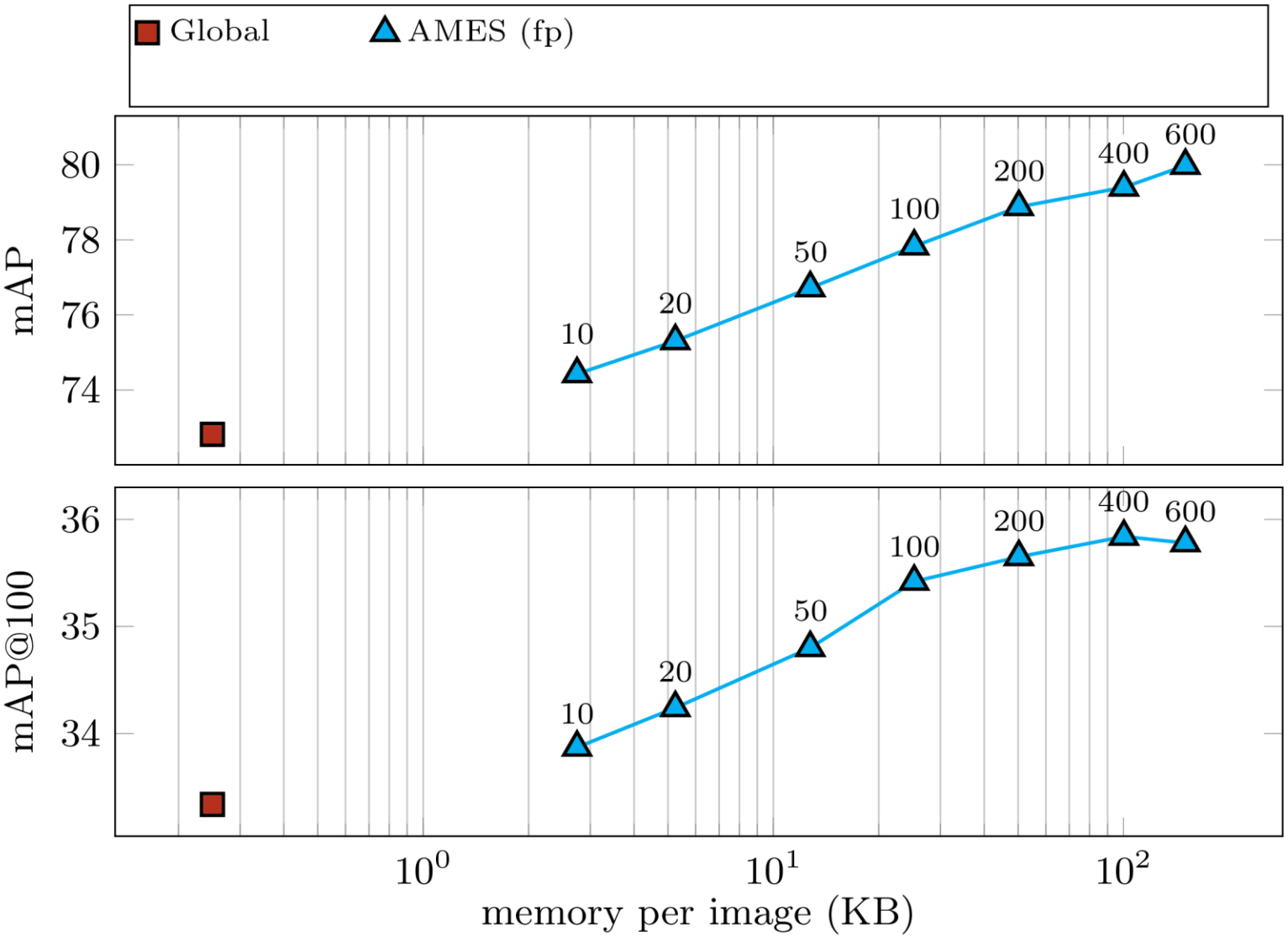
global local

$$\lambda s_g(x, q) + (1 - \lambda) s_l(x, q)$$

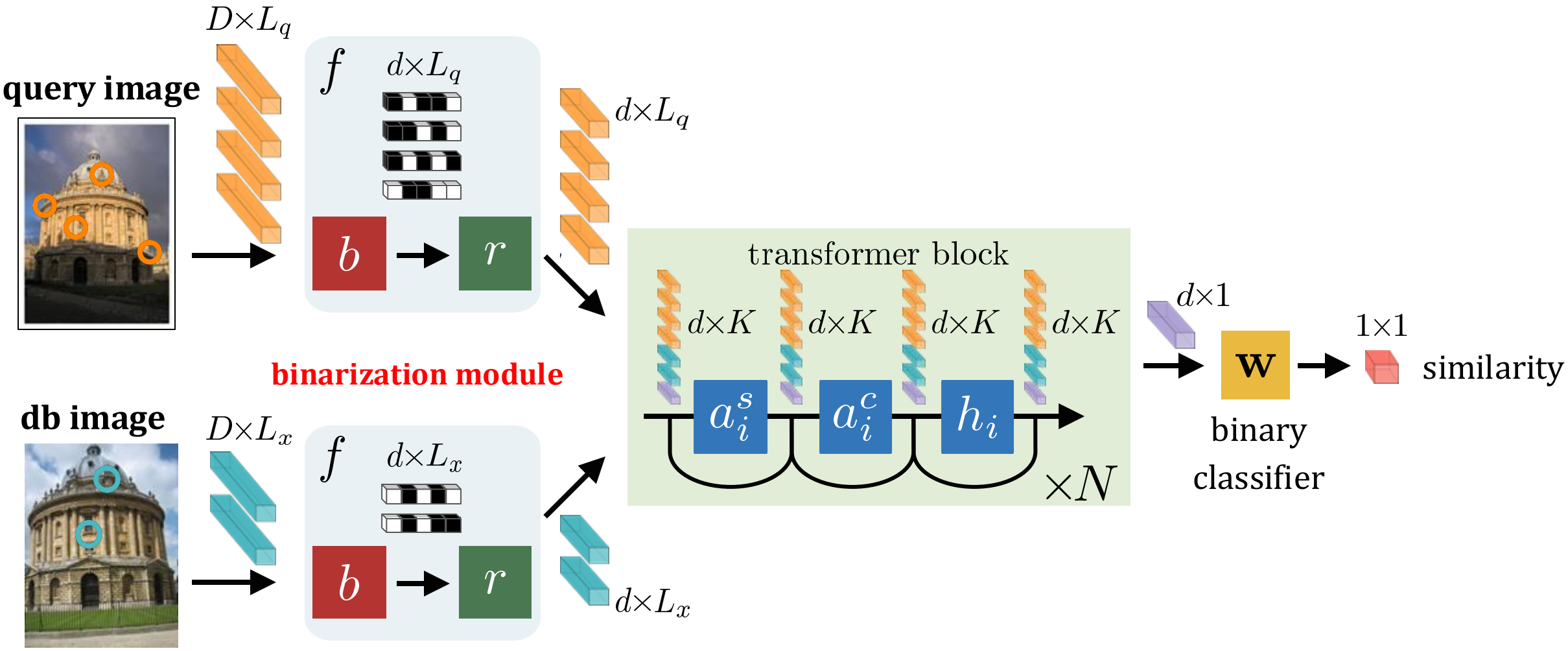
Evaluation on image retrieval

ROP + 1M

GLDv2 test



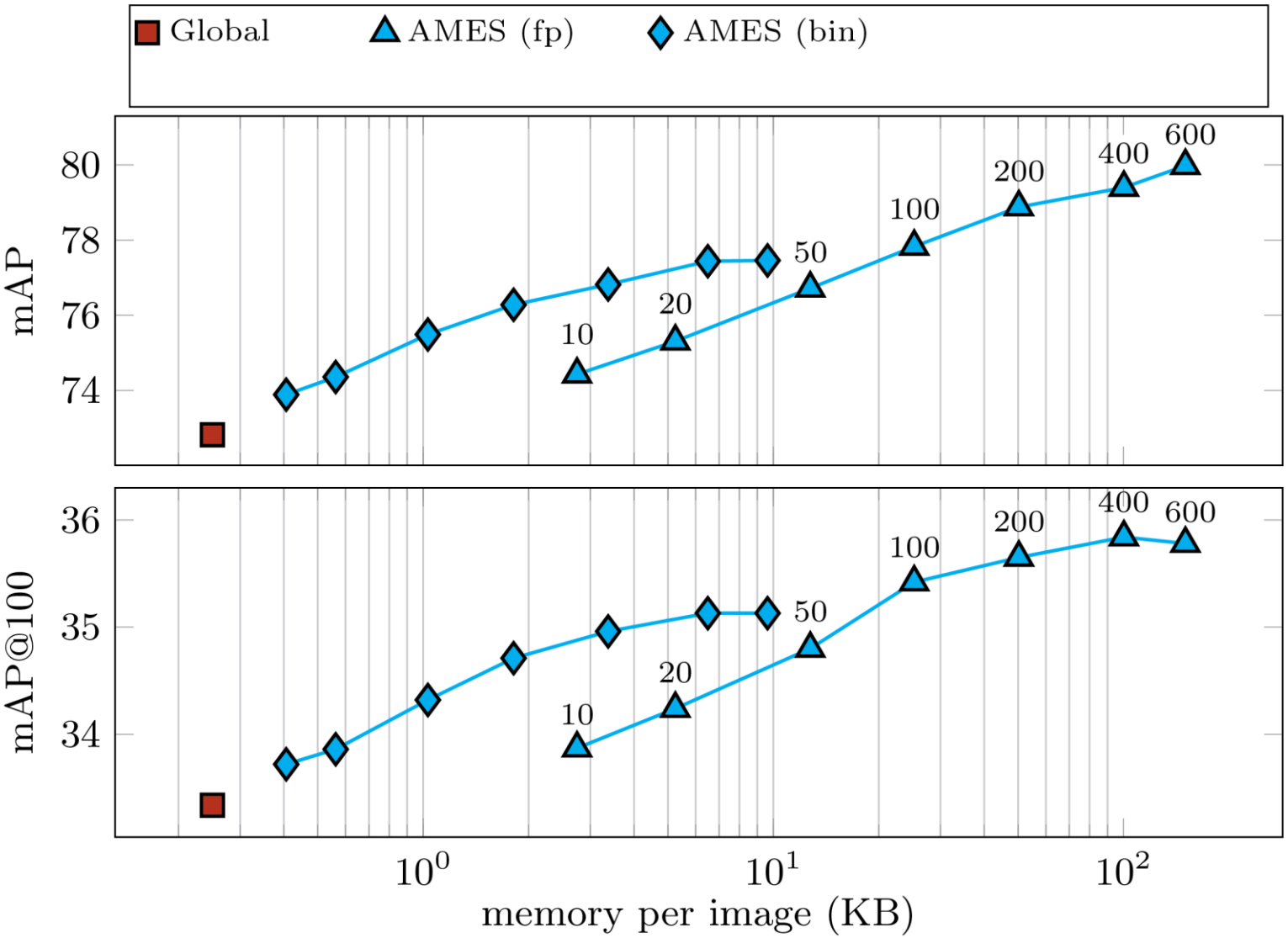
AMES: Asymmetric Memory-Efficient Similarity



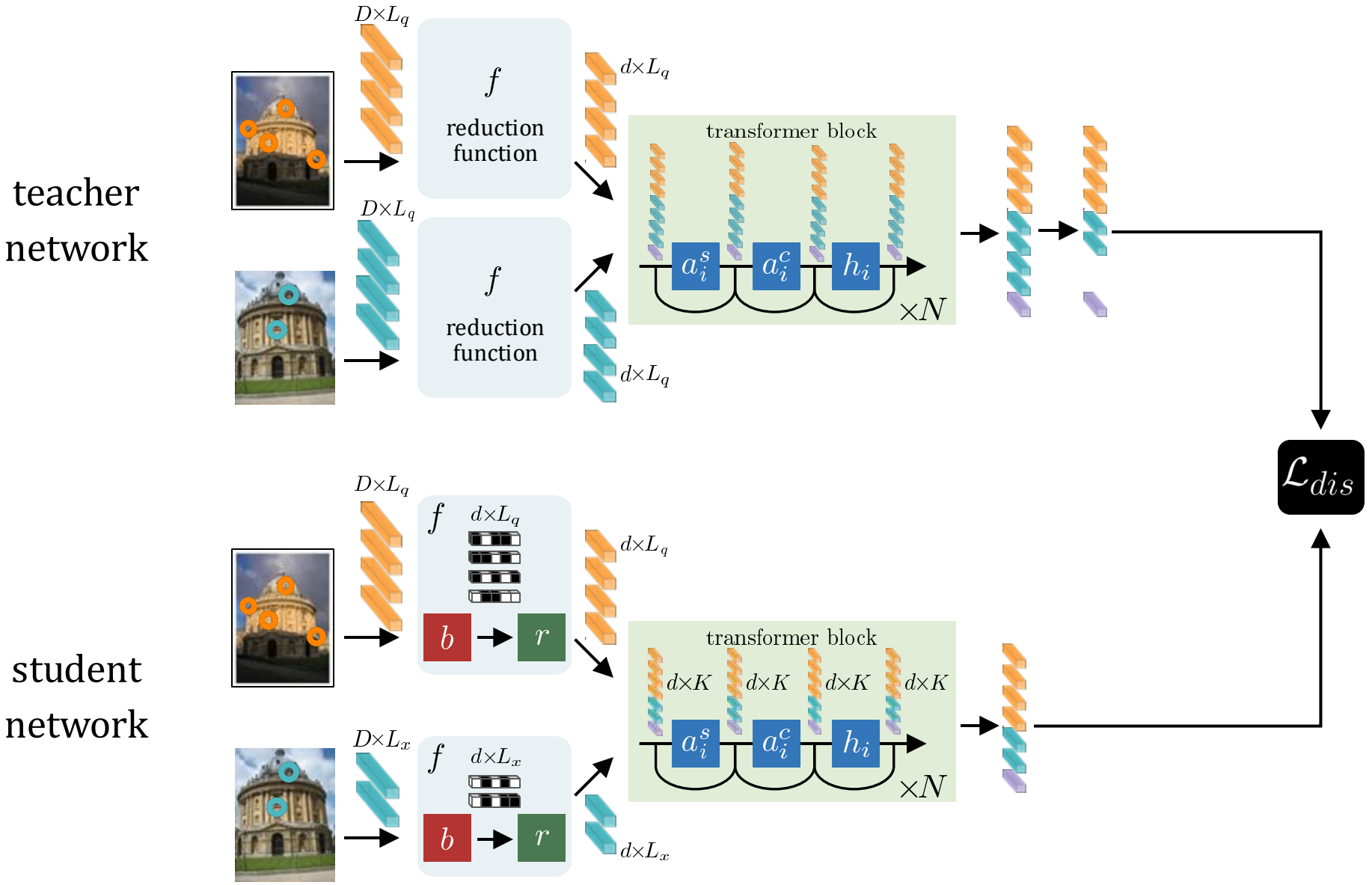
Evaluation on image retrieval

ROP + 1M

GLDv2 test



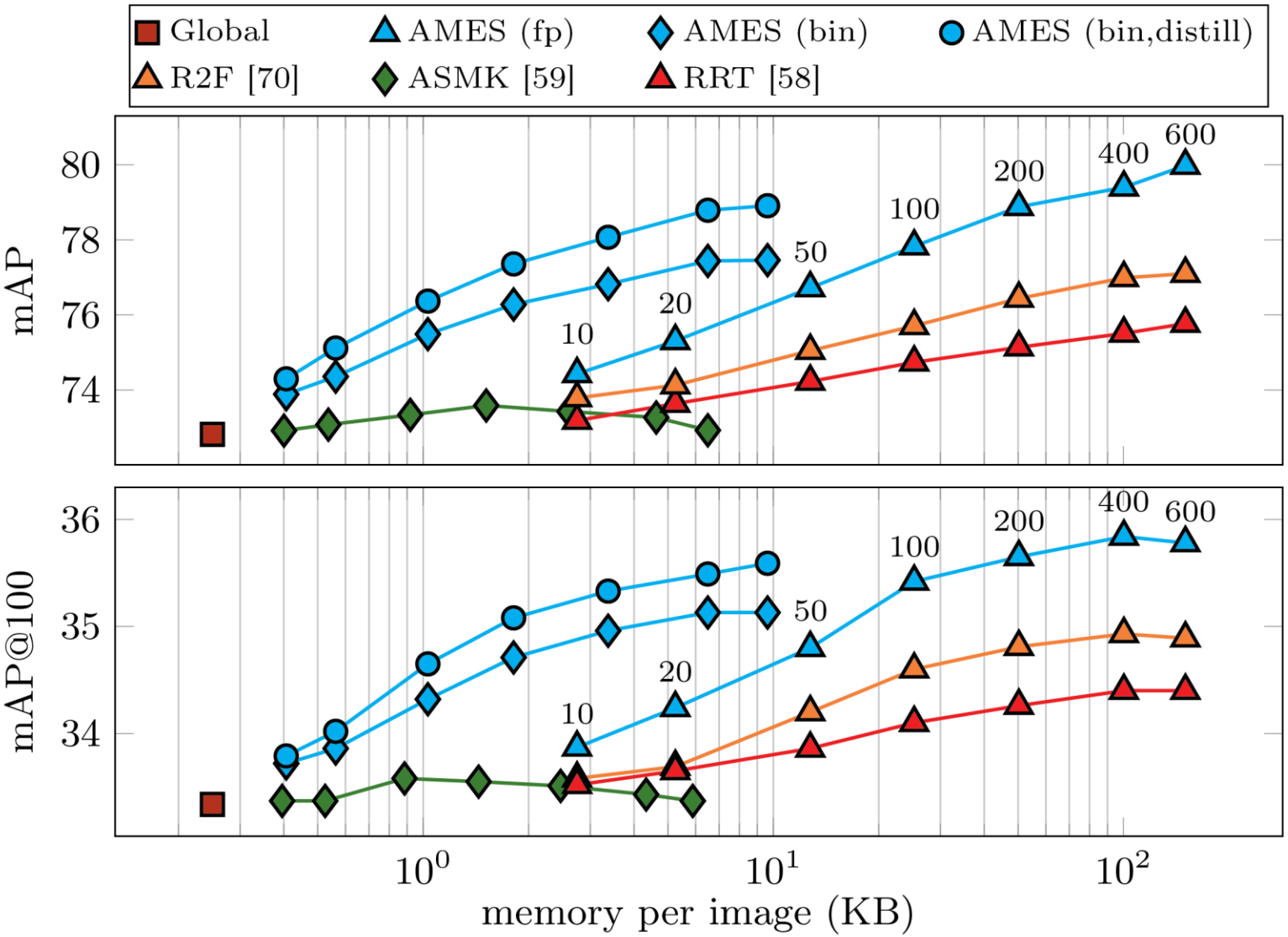
Knowledge distillation



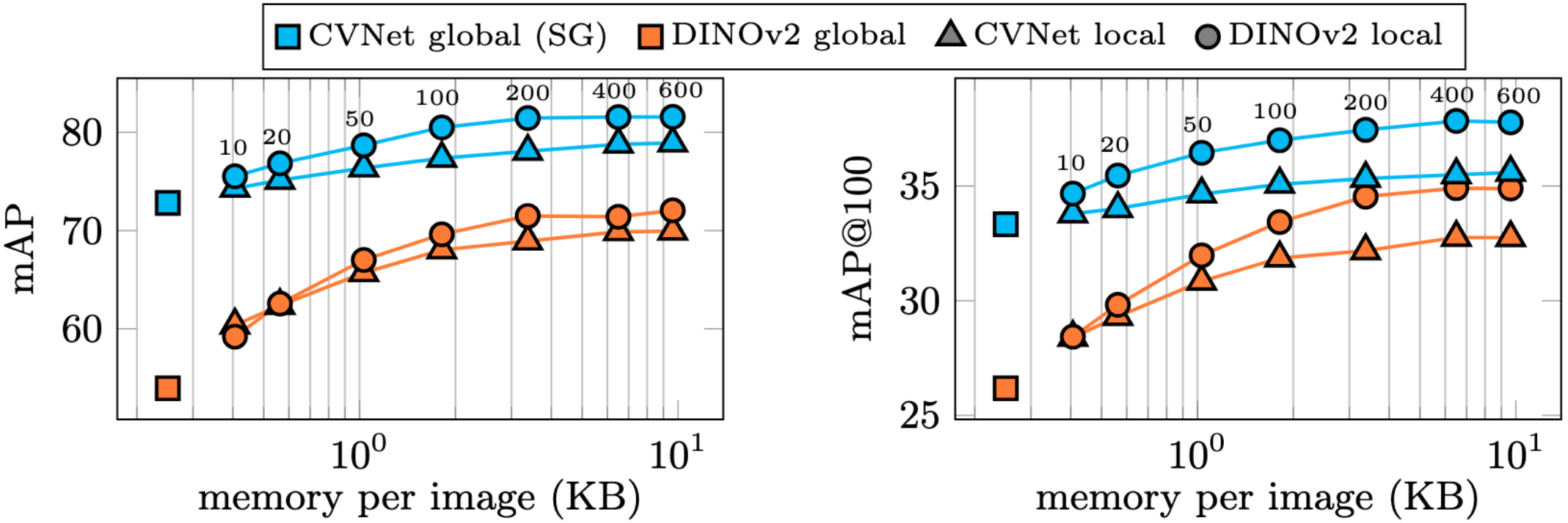
Evaluation on image retrieval

ROP + 1M

GLDv2 test



Network combination

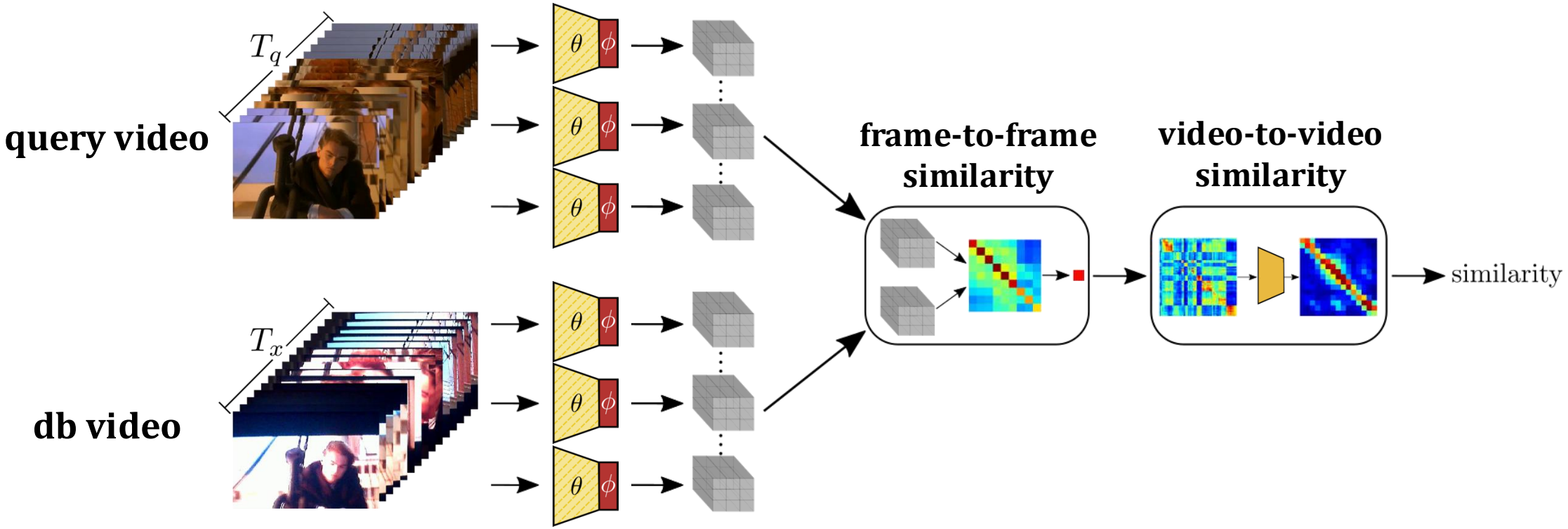


- **DINOv2** provides very good **local descriptors**, while **global** is **not very competitive**
- combining CVNet (SG) global with DINOv2 local achieves **state-of-the-art** performance

Qualitative results

local similarity estimation on videos

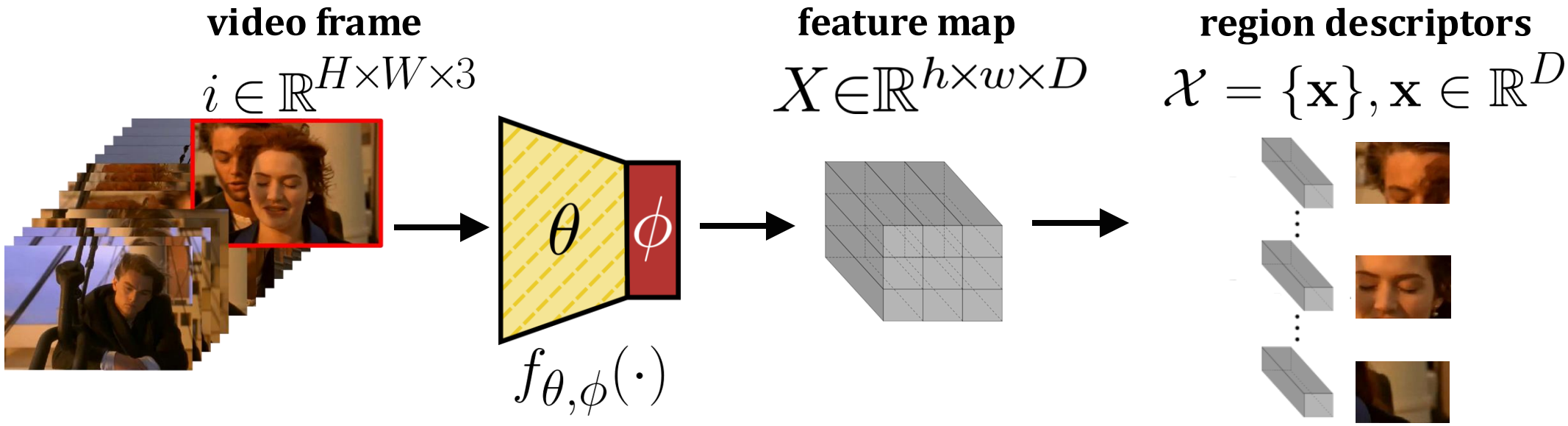
Video Similarity Learning (ViSiL)



- Learn a video similarity function
  - spatial structure of video frames (intra-frame relations)
  - temporal structure of videos (inter-frame relations)

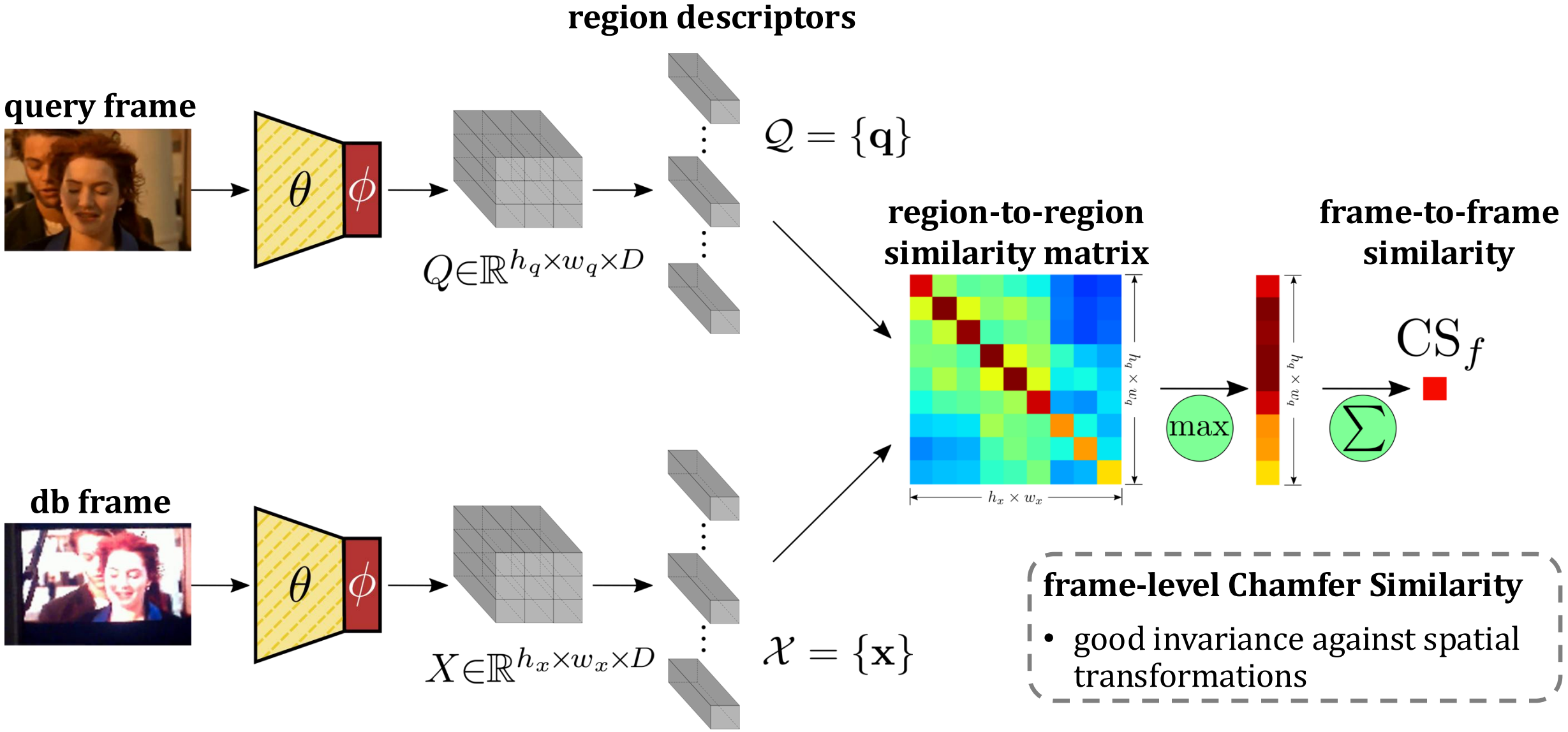
Kordopatis-Zilos et al. "ViSiL: Fine-grained spatio-temporal video similarity learning". ICCV, 2019.

Local descriptor extraction



- extract feature maps from a frozen backbone
- apply region pooling, whitening and attention weighting
- decompose of the feature map into the region vectors

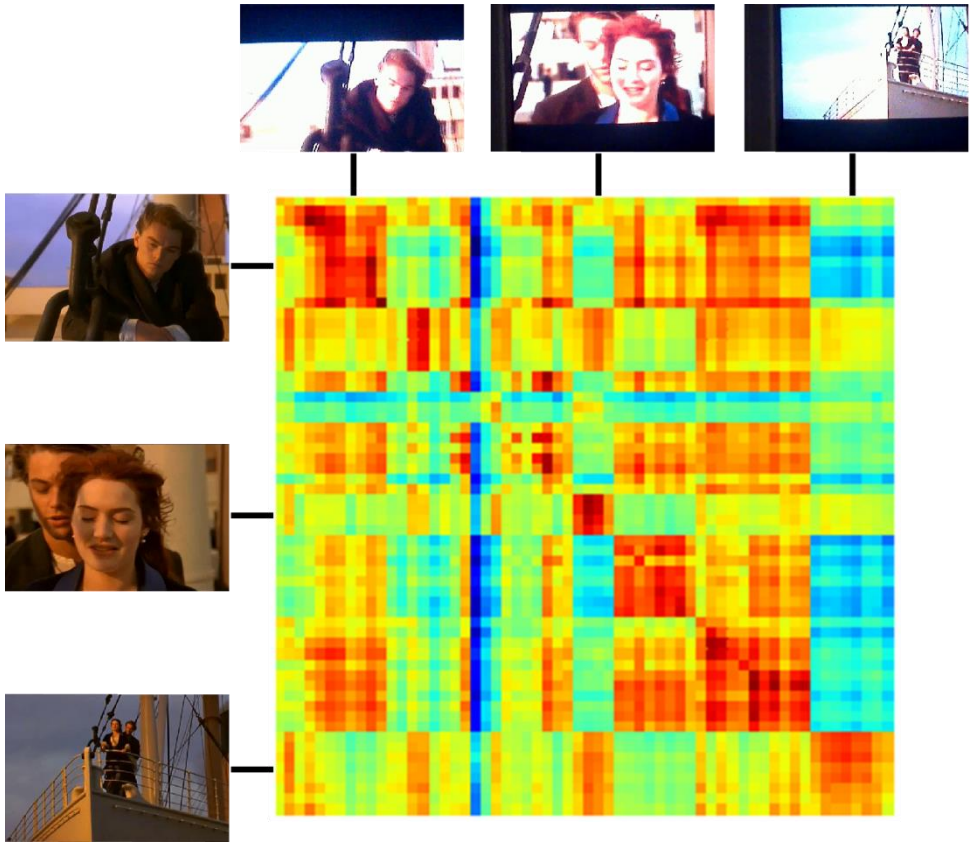
Frame-to-frame similarity



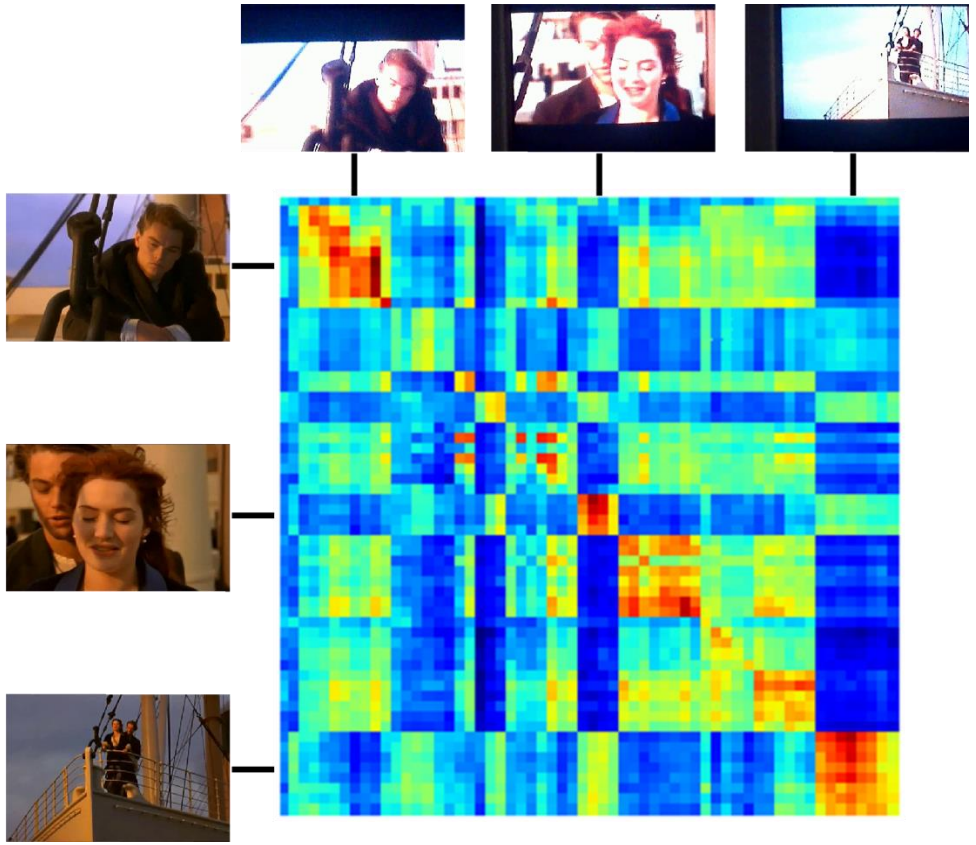
frame-level Chamfer Similarity

- good invariance against spatial transformations

Frame-to-frame similarity matrices

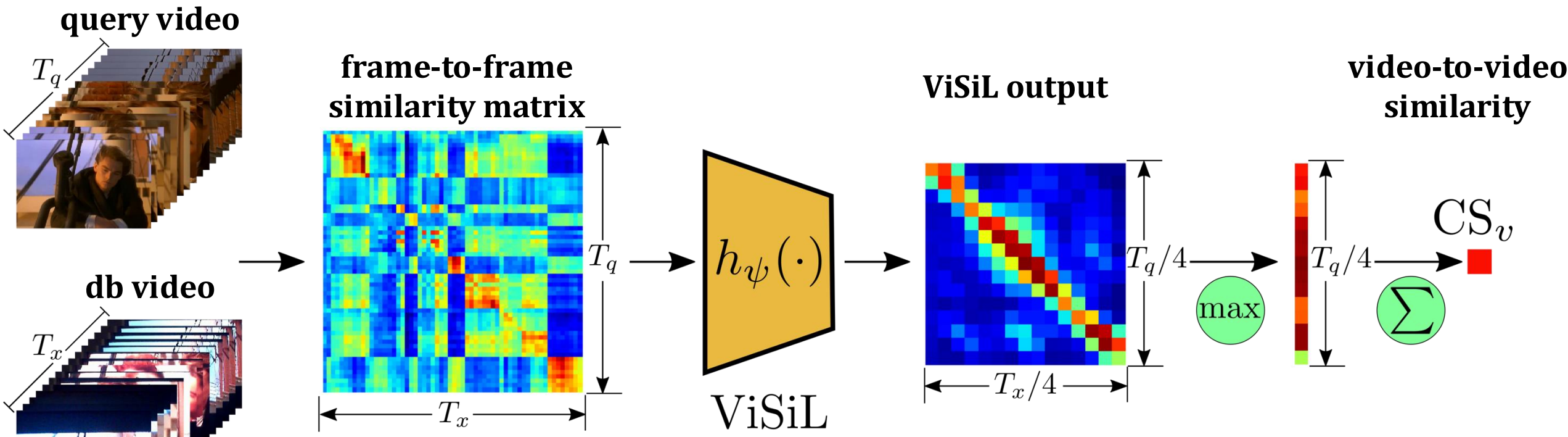


global similarity



local similarity

Video-to-video similarity



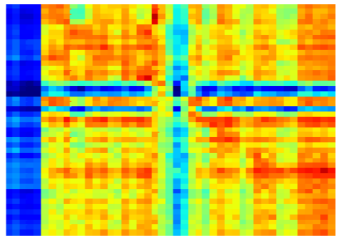
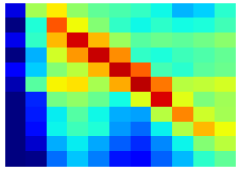
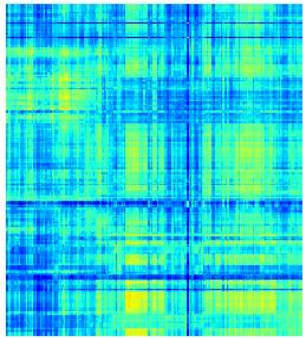
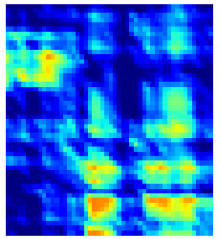
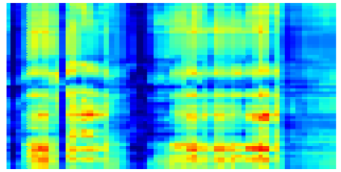
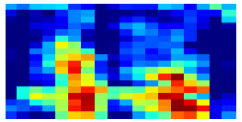
video comparator network

- 4-layer CNN
- captures the **temporal structures** in similarity matrix with the **convolutional filters**

video-level Chamfer Similarity

- good invariance against temporal transformations

Visual examples

	query video	db video	frame-to-frame similarity matrix	ViSiL output
near-duplicate videos				
same event videos				
same action videos				

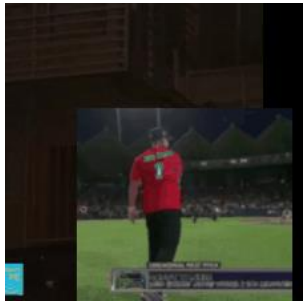
Self-supervised video similarity (S^2VS)

random video

weak augmentations

- conventional geometric transformations
- temporal cropping

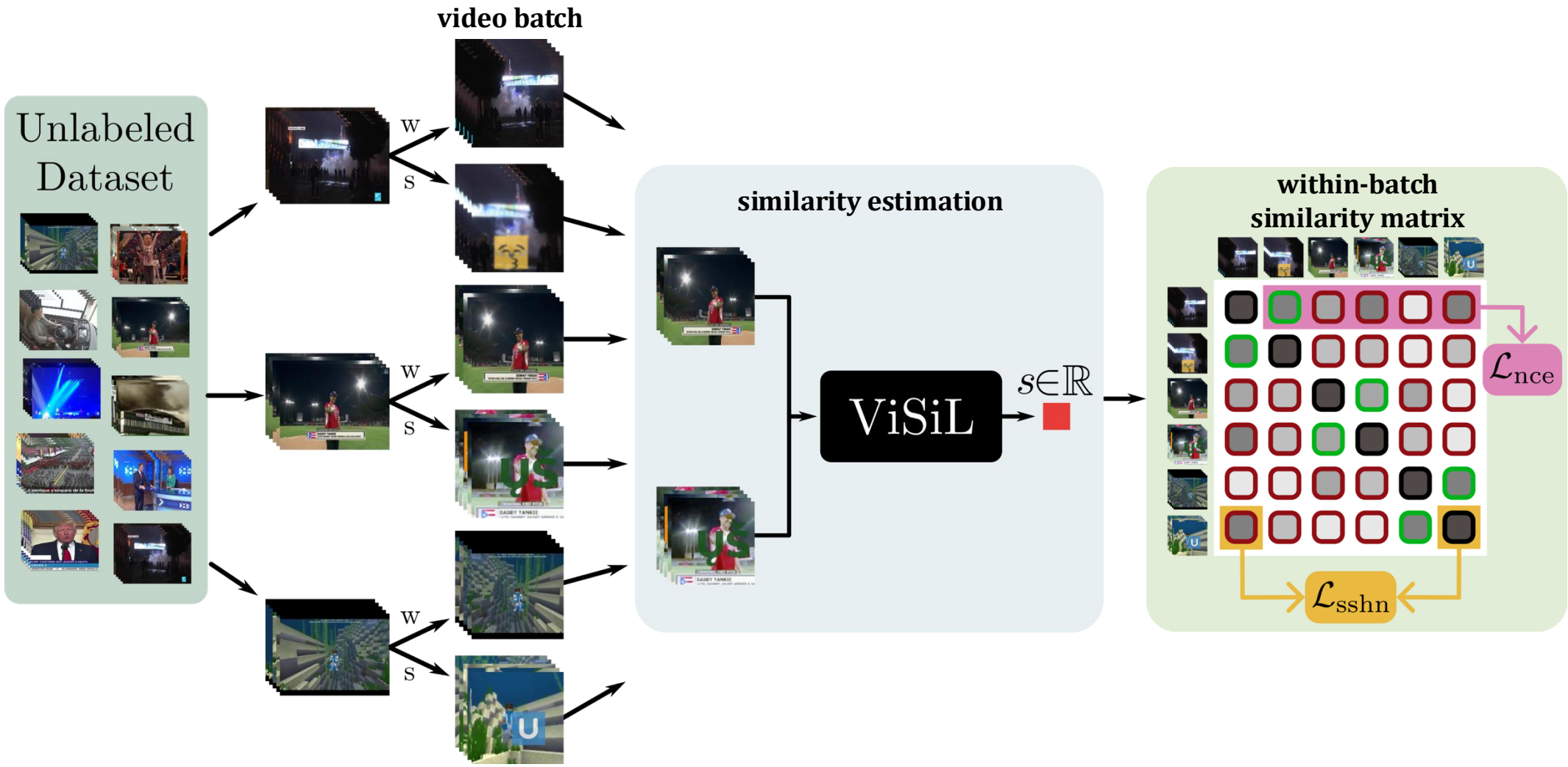
strong augmentations



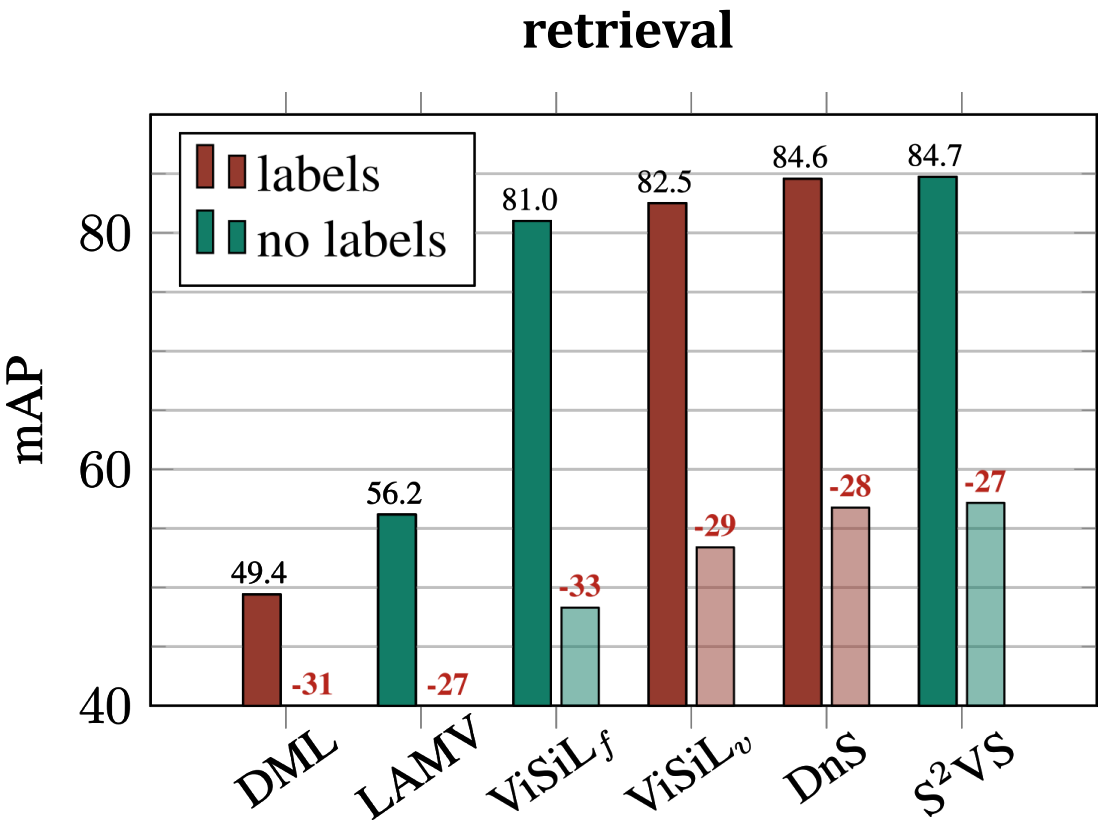
- global transformations
- frame transformations
- temporal transformations
- video-in-video

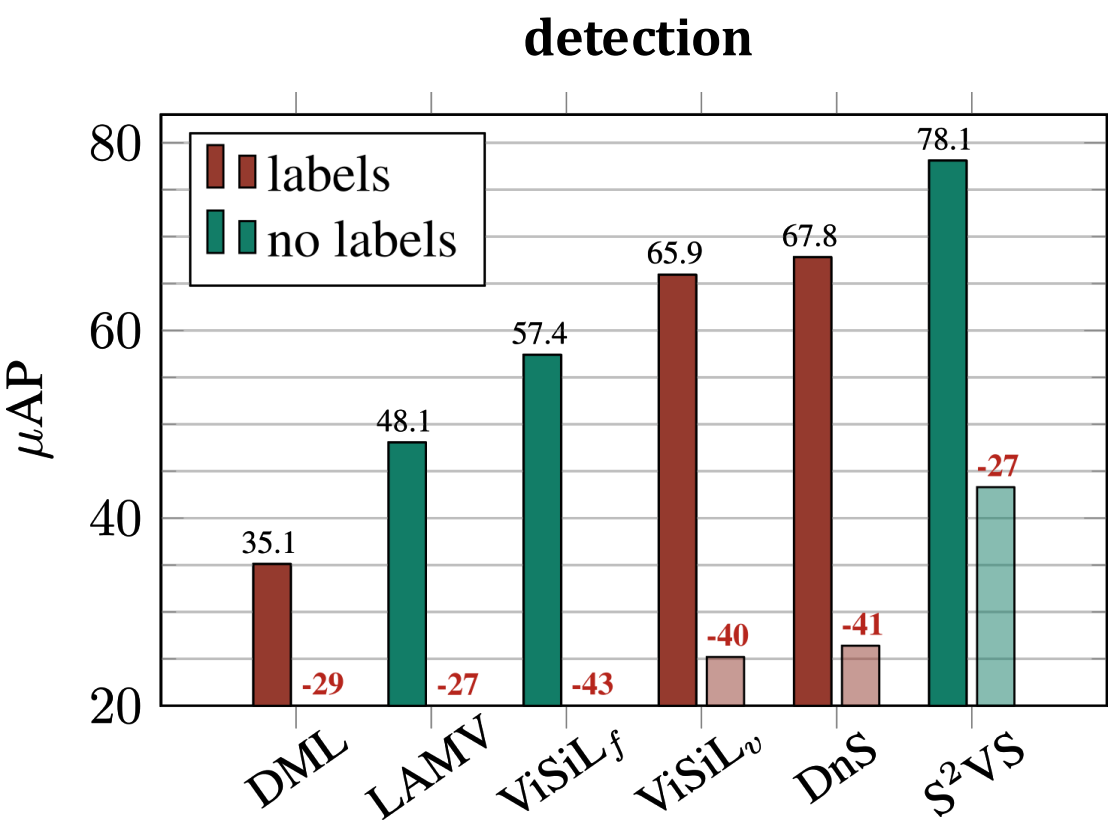
Kordopatis-Zilos et al. “Self-Supervised Video Similarity Learning ." CVPRW, 2023.

ViSiL training w/ self-supervision



Evaluation on FIVR-200K





Take home messages

- Global similarity is a **good scalable solution**
  - **huge space** for improved
- Local similarity is a **very powerful** tool
  - **careful design** to mitigate its limitations

Thank you!

come by poster 178 on *Thursday 10:30-12:30*

Website:

https://gkordo.github.io/

Get in touch:

kordogeo@fel.cvut.cz / @g_kordo

Codes available in:

https://github.com/gkordo