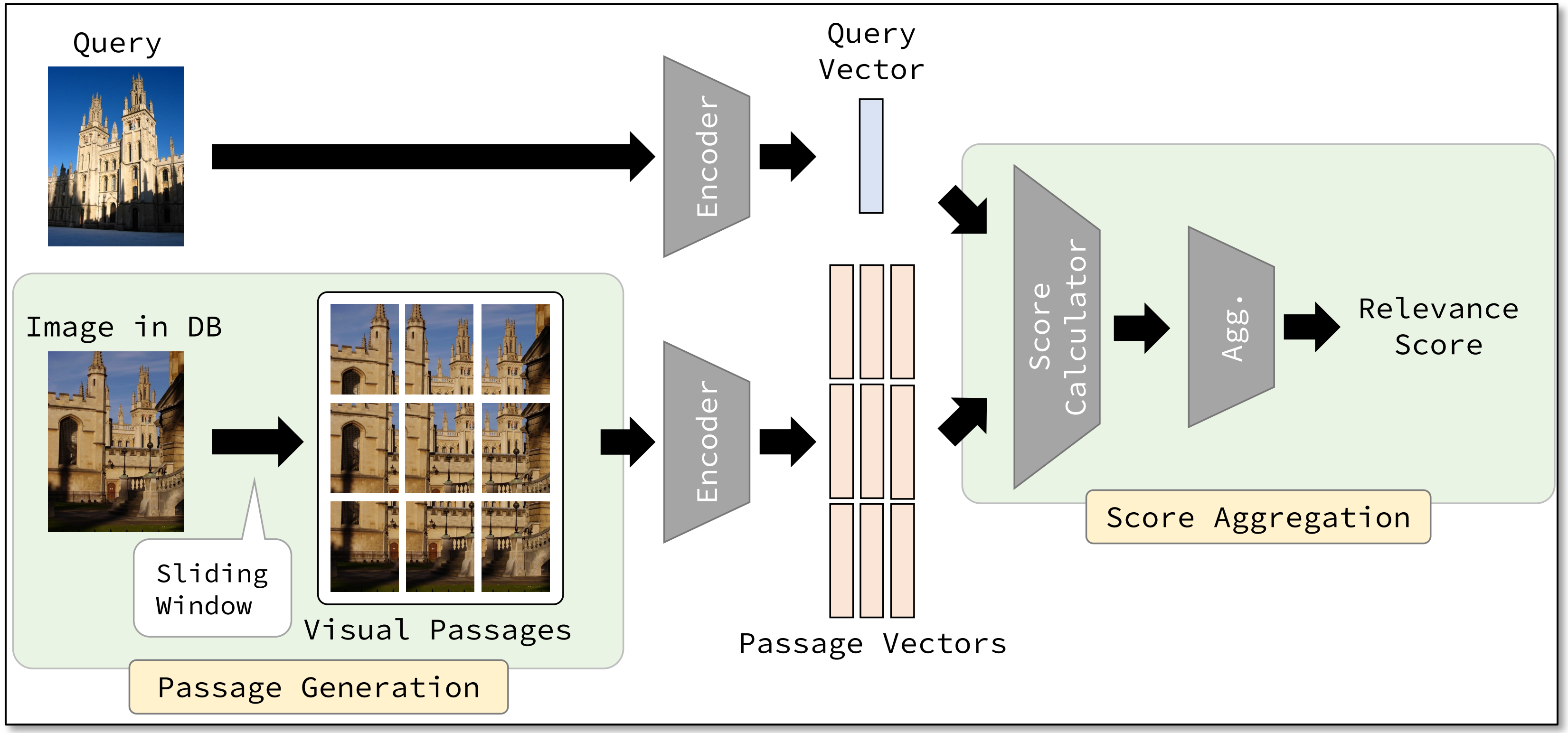
https://taka-coma.pro/pdfs/MIPR2023.pdf

Visual Passage Score Aggregation for Image Retrieval

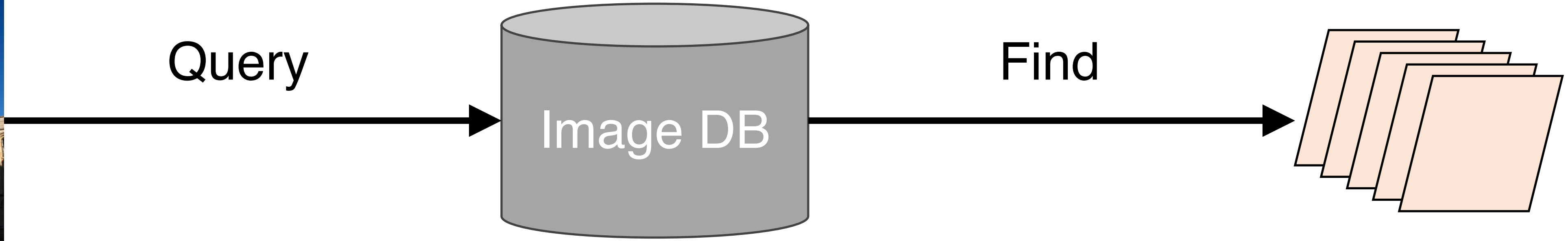
Takahiro Komamizu (taka-coma@acm.org, https://taka-coma.pro/)

Nagoya University



Content-based Image Retrieval

Object-centric image



Images containing that object

- Keys
  - Representations of images (query and images in DB)
    ➔ SIFT, CNN, ViT, etc.
  - Various sizes of objects in each image in DB
    - Some contains an object in the major part of an image.
    - Some contains an object in a small part w/ or w/o occlusion.

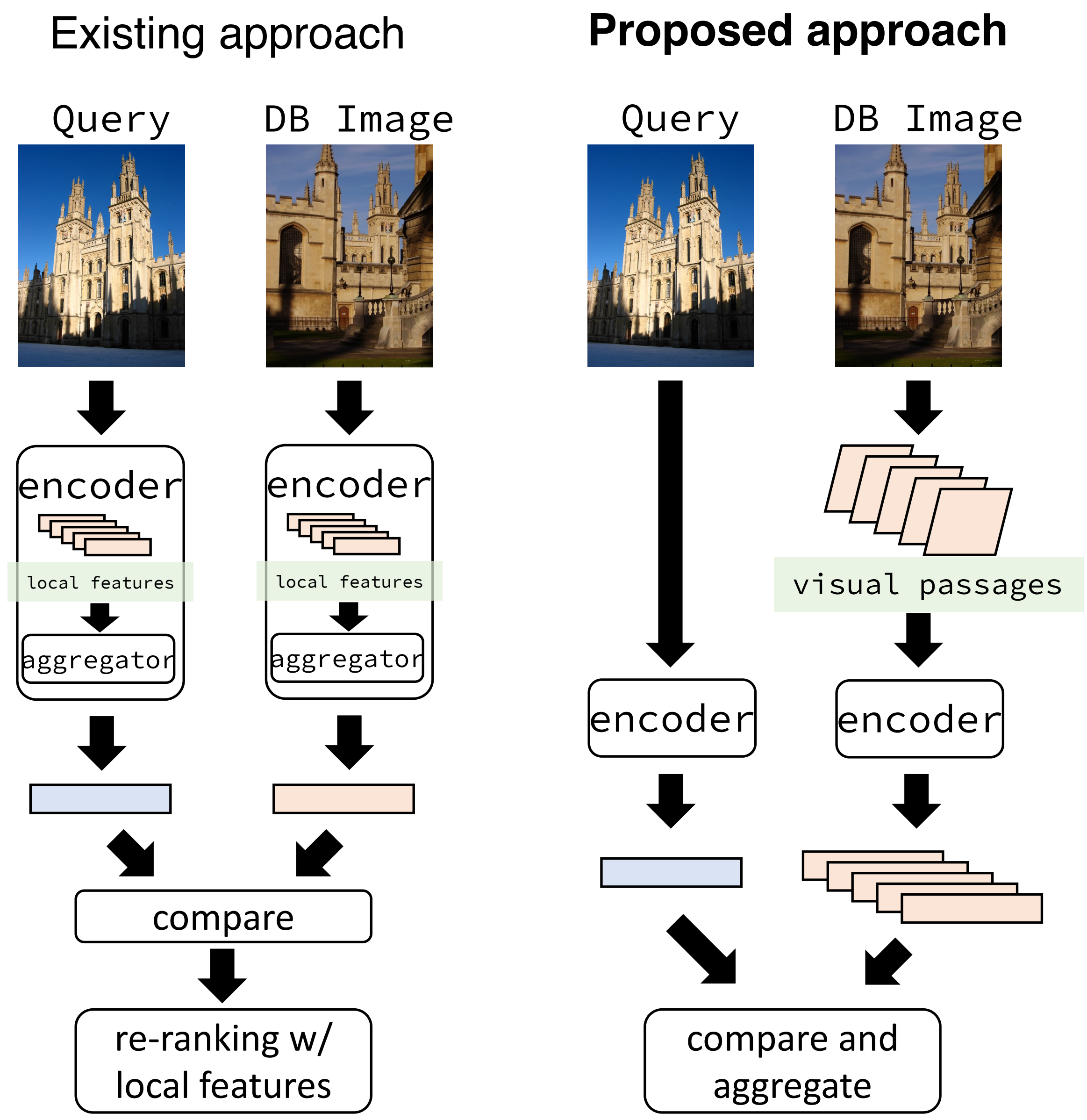
Learning to Rank

- Representations of images may not be good enough for retrieval.
  - k-NN search with the representations is not enough.
- Geometric verification (taking local info more into account)
  - CVNet[5] is the state-of-the-art
  - Find matching of geometric points between query and database images
- Drawback
  - Large amount of training data required
  - Larger inference time
- Common approach
  - Re-ranking is applied for roughly searched top-k images.
    - Compare query image with top-k images (point-wise, pair-wise, and list-wise)

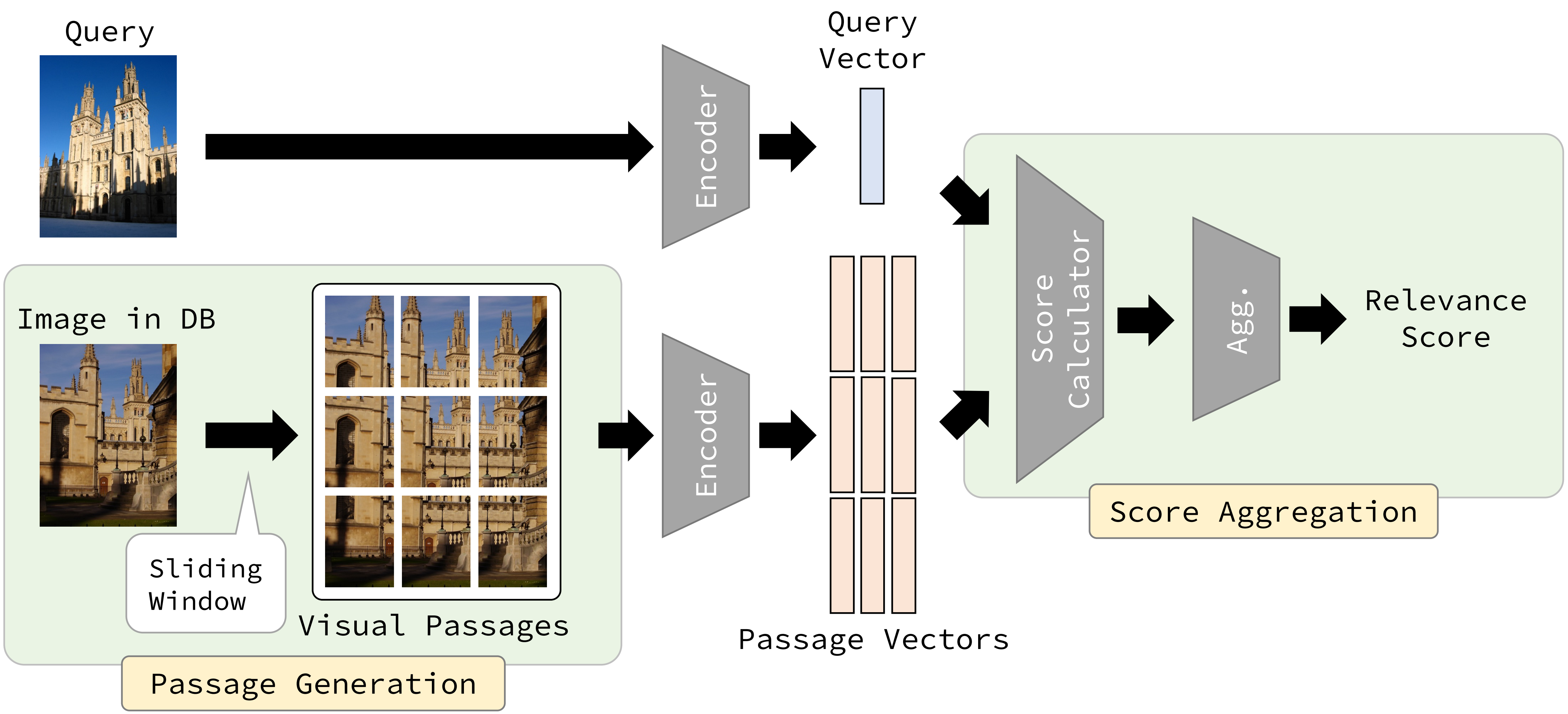
Basic Idea

- A question "*can we realize a single representation to express (complicated) contents of an image*?"
  - ➔ <u>Idea1: Multiple representations for each image.</u>
- The performance of re-ranking approach is bounded by the top-k search results.
  - Expected results not included in the top-k results cannot be re-ranked.

➔ <u>Idea2: Local information into representations of each image</u>

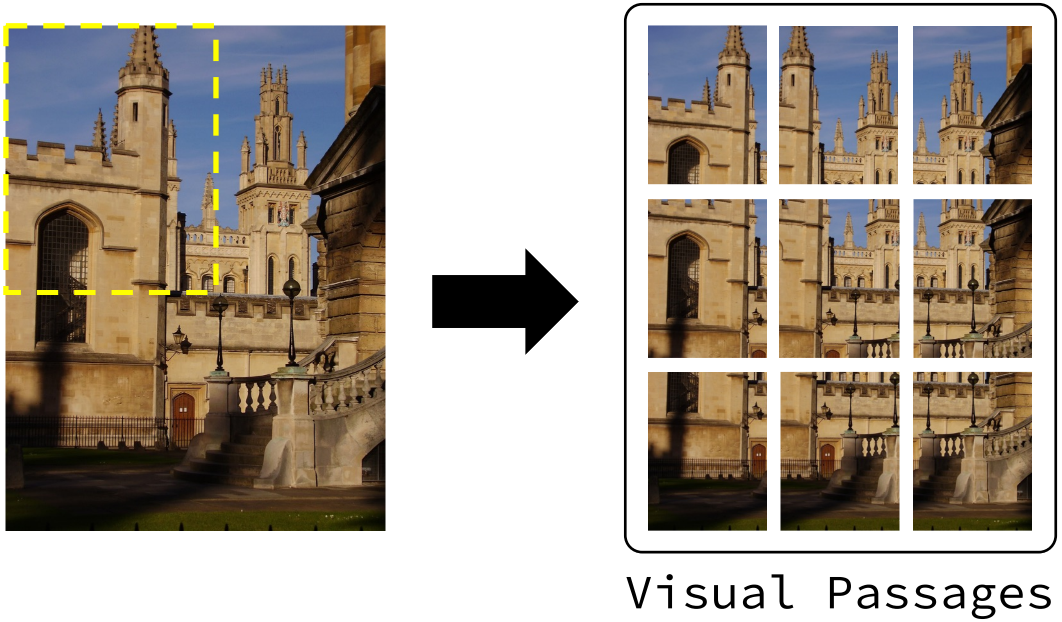


Proposed: Visual Passage Score Aggregation (VPSA)



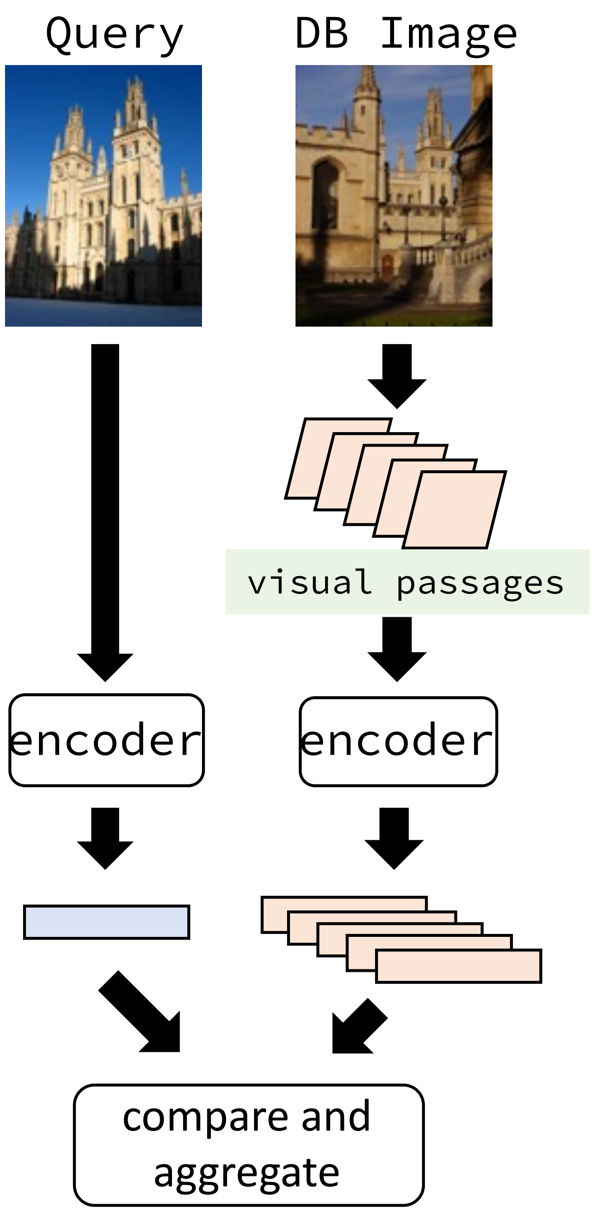
Sliding Window-based Visual Passage Generation

- Visual Passage: a part of image
- Idea in this paper
  - **Coverage**: the set of visual passages covers all part of the image
  - **Overlapping**: not to split objects around the window boundary
  - Same number of visual passages among DB images: to ease the data management



Retrieval using Visual Passages

- Each visual passage is encoded into a vector.
- Retrieval procedure
  - Calculate similarity b/w query and passage
  - For each DB image, aggregate the similarity scores over its visual passages
  - Rank images based on the aggregated scores
- Aggregation strategy: Mean, Max
  - Inspired from text passage-based long document retrieval



Experimental Evaluation

- Dataset: Revisited Oxford5K / Paris6K + Destructor set (1M)
  - Images about buildings, destructor set contains confusing images
  - 70 queries for each
- Metrics: MAP (mean average precision)
- Comparative methods
  - **NN**: Nearest neighbor method (baseline)
  - **DOLG**[32], **TBR**[29]: Local feature aggregation approaches
  - **DFS** (Offline Diffusion)[31]: an efficient diffusion-based approach
  - **RRT**[23], **CVNet**[13]: Re-ranking approaches

Aggregation Functions

- Max was the best.
  - Local features via visual passages increased the retrieval performance.
  - The most similar part of an image is important when the target objects appeared differently in DB images.

Method		MEDIUM $\mathcal{R}$Oxf	MEDIUM $\mathcal{R}$Par	HARD $\mathcal{R}$Oxf	HARD $\mathcal{R}$Par
Baseline:	NN	80.2	90.3	63.1	79.1
Proposed:	VPSA-Mean	71.5	87.6	42.7	73.3
Proposed:	VPSA-Max	**85.5**	**91.2**	**70.6**	**81.6**

- Mean was worse than NN, and its performance drop in HARD datasets was larger.
  - Treating all passages equally had negative effect.
    ➔ Weighted approach can be a future solution.

Comparison to Comparative Methods

Method	Base Feature	Approach	Medium $\mathcal{R}$Oxf	Medium +$\mathcal{R}$1M	Medium $\mathcal{R}$Par	Medium +$\mathcal{R}$1M	Hard $\mathcal{R}$Oxf	Hard +$\mathcal{R}$1M	Hard $\mathcal{R}$Par	Hard +$\mathcal{R}$1M
DOLG [32]	R101-GLDv2-clean	LF	81.5	77.4	91.0	83.3	61.1	54.8	80.3	66.7
TBR [29]	R101-GLDv2-clean	LF	82.3	70.5	89.3	76.7	66.6	47.3	78.6	55.9
DFS (10^3) [31]	R101-CVNet-Global	DFS	78.6	76.0	90.9	<u>88.5</u>	59.8	57.3	<u>83.8</u>	<u>79.5</u>
RRT [23] (top100)	R50-GLDv2-clean	RR	78.1	67.0	86.7	69.8	60.2	44.1	75.1	49.4
RRT [23] (top400)	R50-GLDv2-clean	RR	80.5	70.6	89.1	73.8	64.2	49.5	78.1	55.6
CVNet [13] w/o RR	R101-CVNet-Global	NN	80.2	74.0	90.3	80.6	63.1	53.7	79.1	62.2
CVNet [13] (top100)	R101-CVNet-Global	RR	85.6	79.6	90.6	81.5	72.9	64.5	80.4	66.2
CVNet [13] (top400)	R101-CVNet-Global	RR	<u>**87.2**</u>	<u>**81.9**</u>	<u>91.2</u>	83.8	<u>**75.9**</u>	<u>**67.4**</u>	81.1	69.3
VPSA-Max	R101-CVNet-Global	VP	85.5	79.0	91.2	81.3	70.6	60.5	81.6	63.3
VPSA-Max + DFS (10^3)	R101-CVNet-Global	VP+DFS	<u>85.6</u>	<u>81.2</u>	<u>**92.6**</u>	<u>**89.6**</u>	<u>72.7</u>	<u>64.7</u>	<u>**86.5**</u>	<u>**80.1**</u>

- VPSA-Max performed superior to the most of methods, and was comparable with CVNet (the state-of-the-art).
- To combine the diffusion mechanism, the performance increased.

Efficiency

- Though the retrieval performance was comparable to CVNet, retrieval time of VPSA was smaller.
  - Re-ranking methods were still challenging in the efficient inference.
- VPSA took larger time than NN.
  - The number of vectors stored in a database can be easily large.

Model	Time (70 queries)	Time per Query
NN	0.62 sec	0.009 sec
VPSA-Max	0.94 sec	0.013 sec
VPSA-Max + DFS	28.12 sec	0.402 sec
CVNet (top100)	9 min 25 sec	8.071 sec
CVNet (top400)	28 min 53 sec	24.757 sec

Conclusion and Future Work

- Conclusion
  - VPSA: Visual Passage Score Aggregation
    - Visual passage: a crop of an image
    - Aggregation: similarity scores are aggregated via Max or Mean function
  - Experiment showed the effectiveness and efficiency of VPSA
- Future Work
  - To explore methods to improve effectiveness, other representation schemes (like ViT and Swin Transformer) will be tested.
  - To seek a way of combining strengths of VPSA and re-ranking methods.