Large-Scale Image Retrieval with Attentive Deep Local Features

ICCV 2017

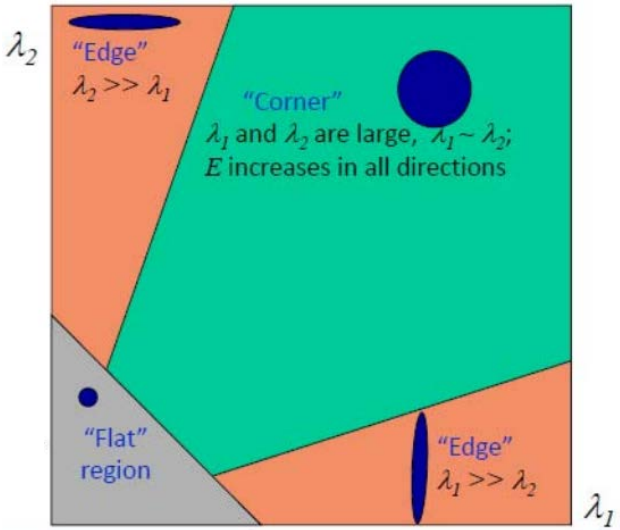
2021.04.27

Sebin Lee

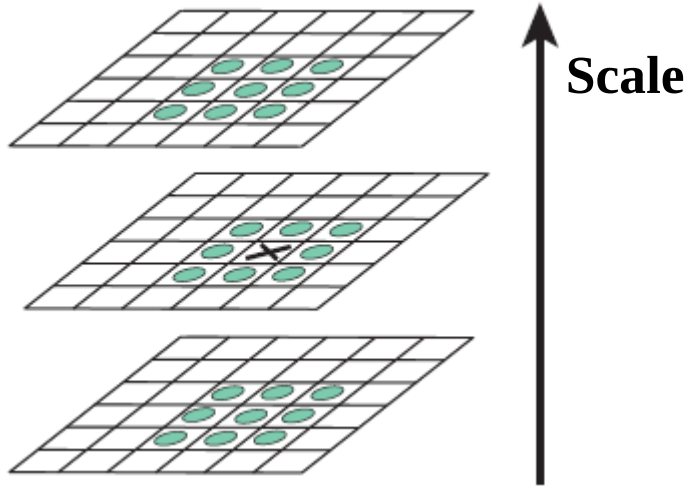
Contents

Recap: Keypoint

- Important point of image (e.g., corner)
- The keypoint itself is not useful for image retrieval.
- e.g., Harris corner detector, Local maxima of DoG



< Harris corner detector >

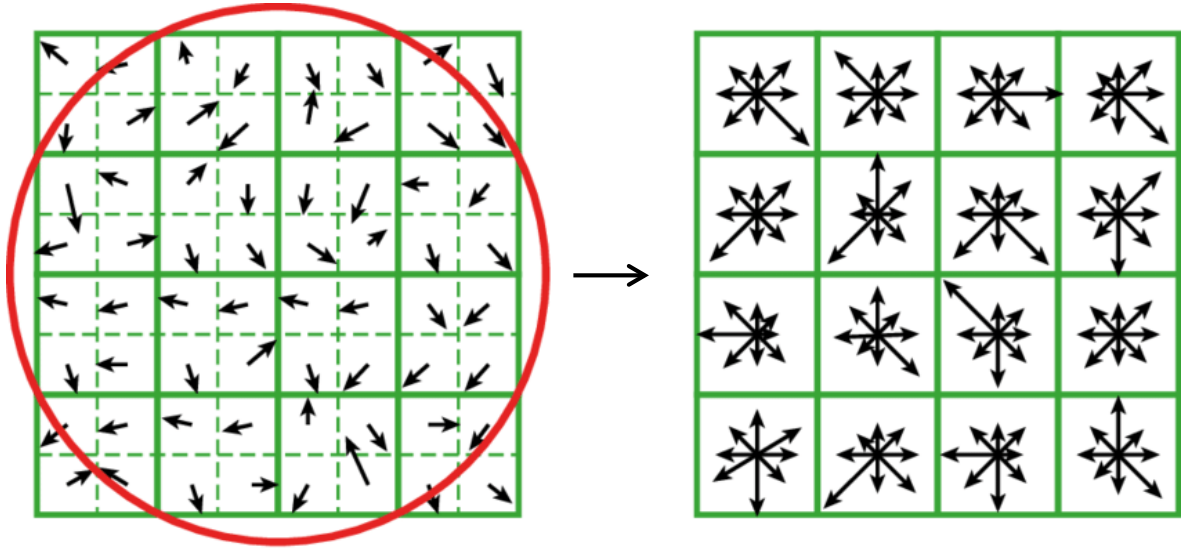


< Local maxima of DoG >

picture from Sung-eui Yoon slide, originally made by Kristen Grauman and David Lowe

Recap: Local Descriptor

- The local descriptor is used because the keypoint itself is not useful.
- The local descriptor means a compact representation of the image patch centered on the extracted keypoint.
- e.g., SIFT



< Gradients of image patch > < SIFT descriptor >

Classical Keypoint & Local Descriptor Properties

- **Repeatable**
- **Distinctive**
- **Invariant**
- **Adequate Quantity**

} **Sparse keypoint & descriptor**

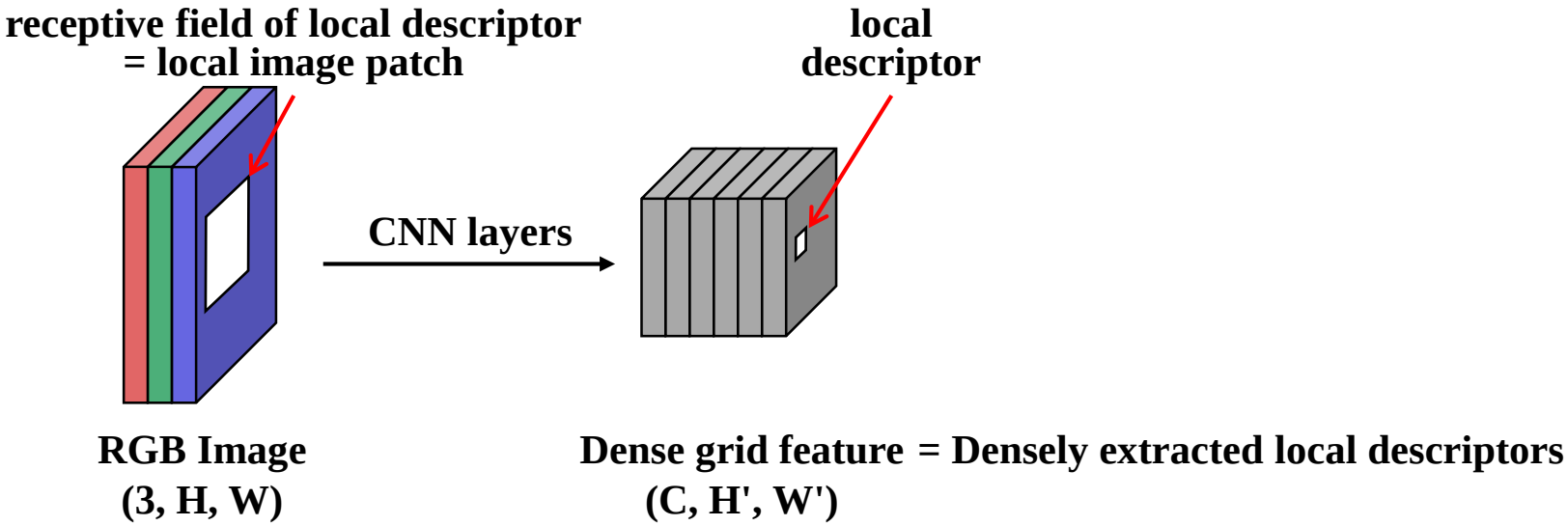
< Keypoints of Image>

How about deep feature?

picture from Sung-eui Yoon slide, originally made by Denis Simakov

Deep Feature

- CNNs generate uniformly dense grid feature map.
- We can regard the dense grid feature map as a grid of local descriptors.



Motivation

- **Unlike classical methods, the CNNs generate uniformly dense grid local descriptors.**
- **Therefore, local descriptors are extracted from regions that have no value as keypoints. (e.g., texture-less region)**
- **Many local descriptors containing unnecessary ones hinder search and making codebook.**

< Query feature contains unnecessary local descriptors (people) >

∴ We needs keypoint selection that selects only helpful keypoints for efficient and accurate image retrieval.

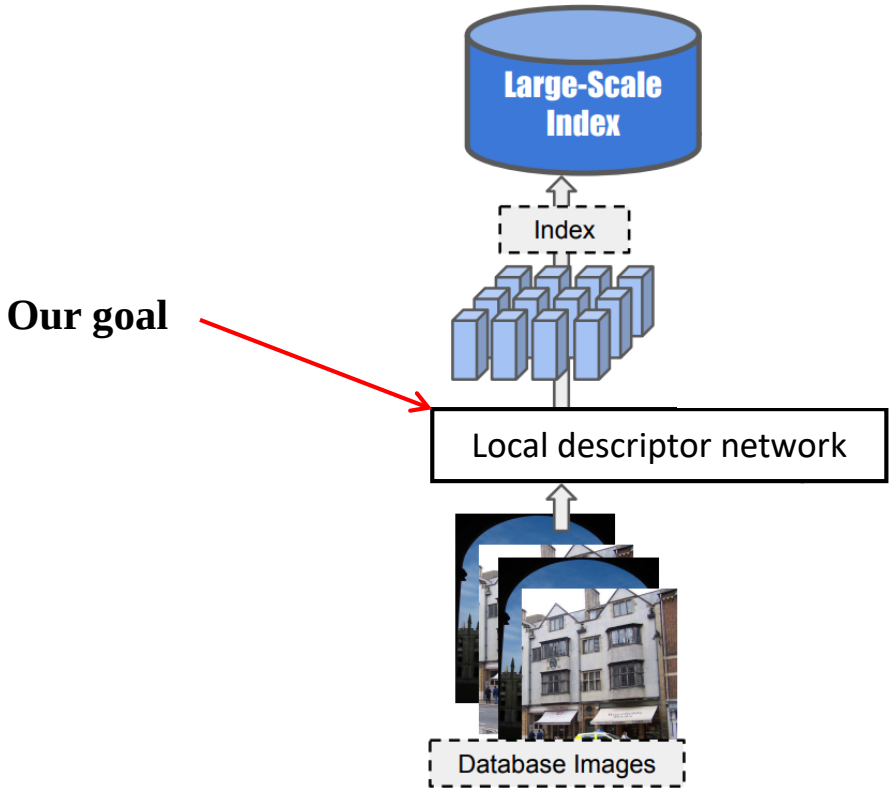
picture from Sung-eui Yoon slide, originally made by Denis Simakov

Contents

- **Background & Motivation**
- **Our Approach**
- **Results**
- **Summary**
- **Quiz**

Goal

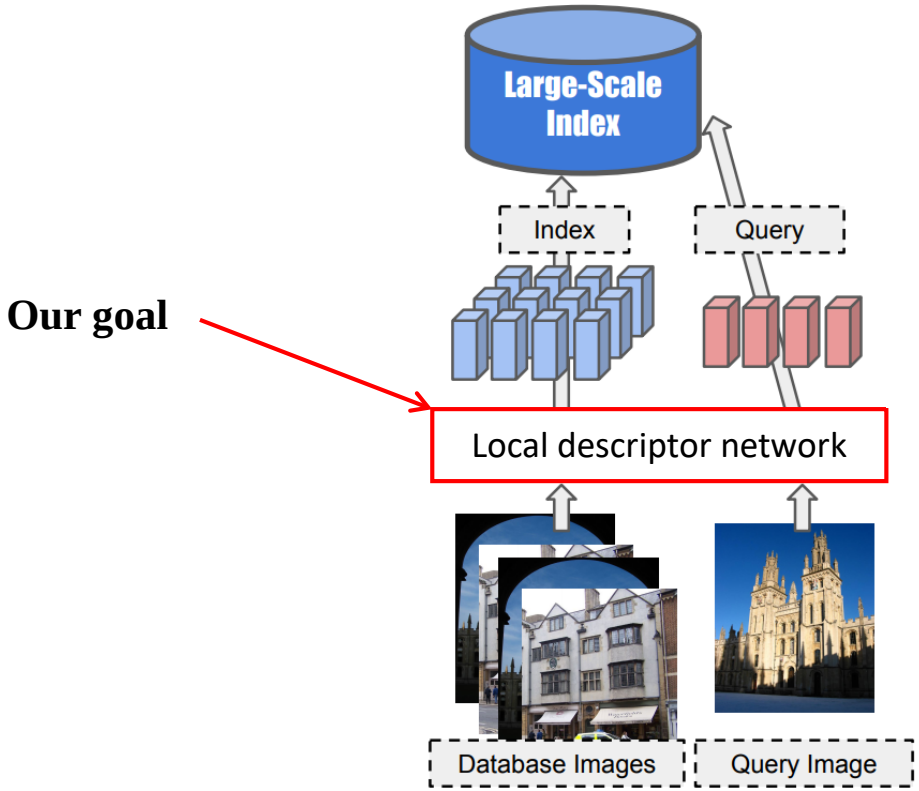
- **Train a local descriptor network using keypoint selection for efficient and accurate image retrieval.**



< Overall Pipeline of Image Retrieval >

Goal

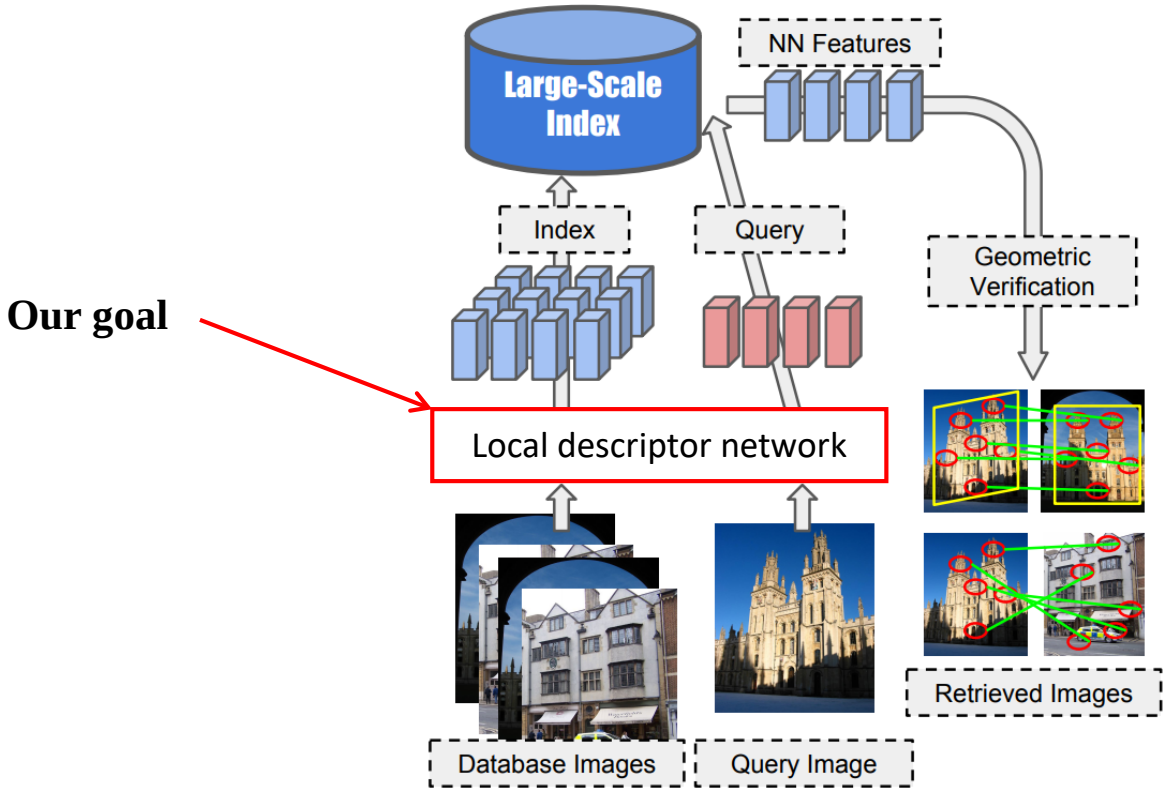
- **Train a local descriptor network using keypoint selection for efficient and accurate image retrieval.**



< Overall Pipeline of Image Retreival >

Goal

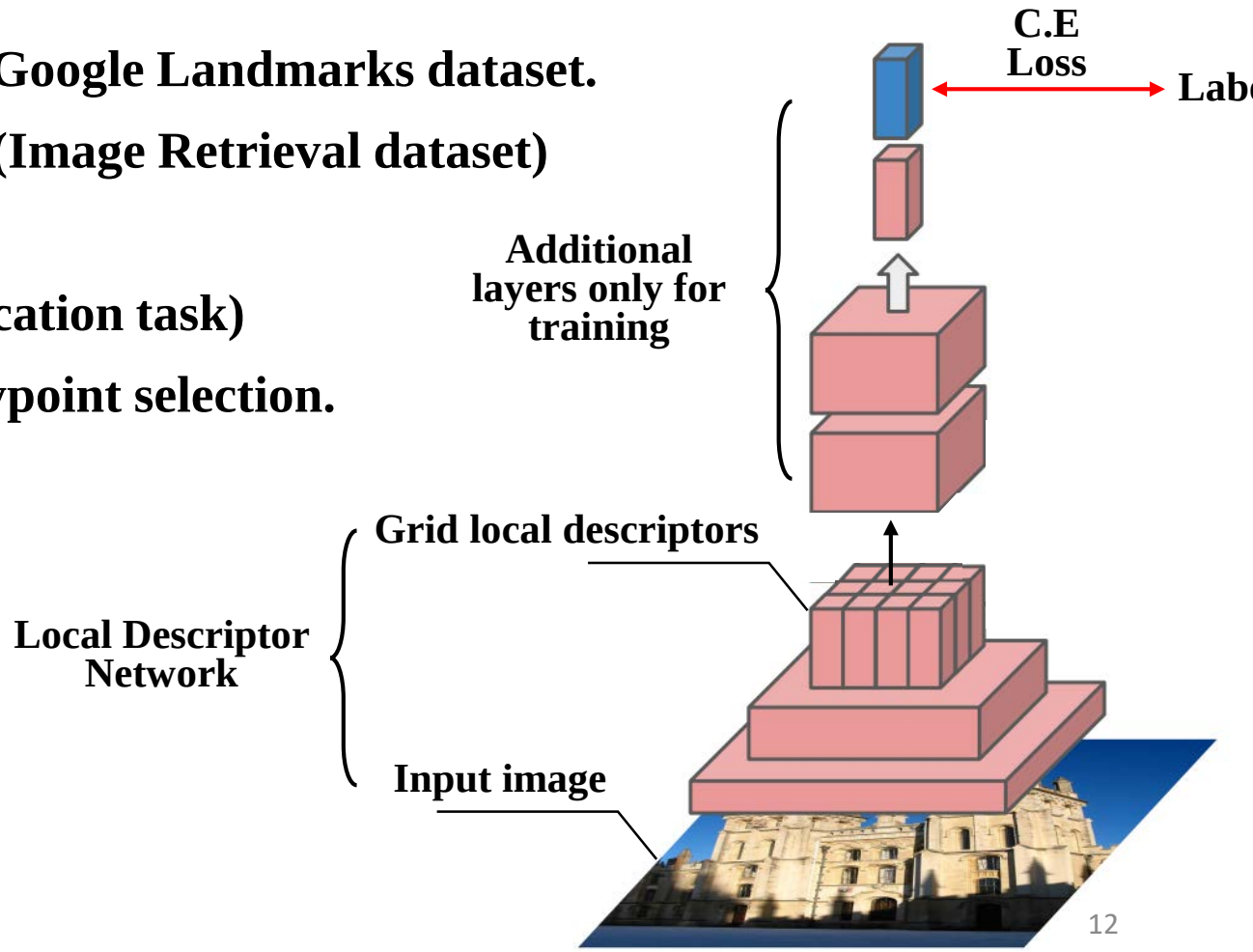
- **Train a local descriptor network using keypoint selection for efficient and accurate image retrieval.**



< Overall Pipeline of Image Retreival >

Train Local Descriptor Network by using Classification task

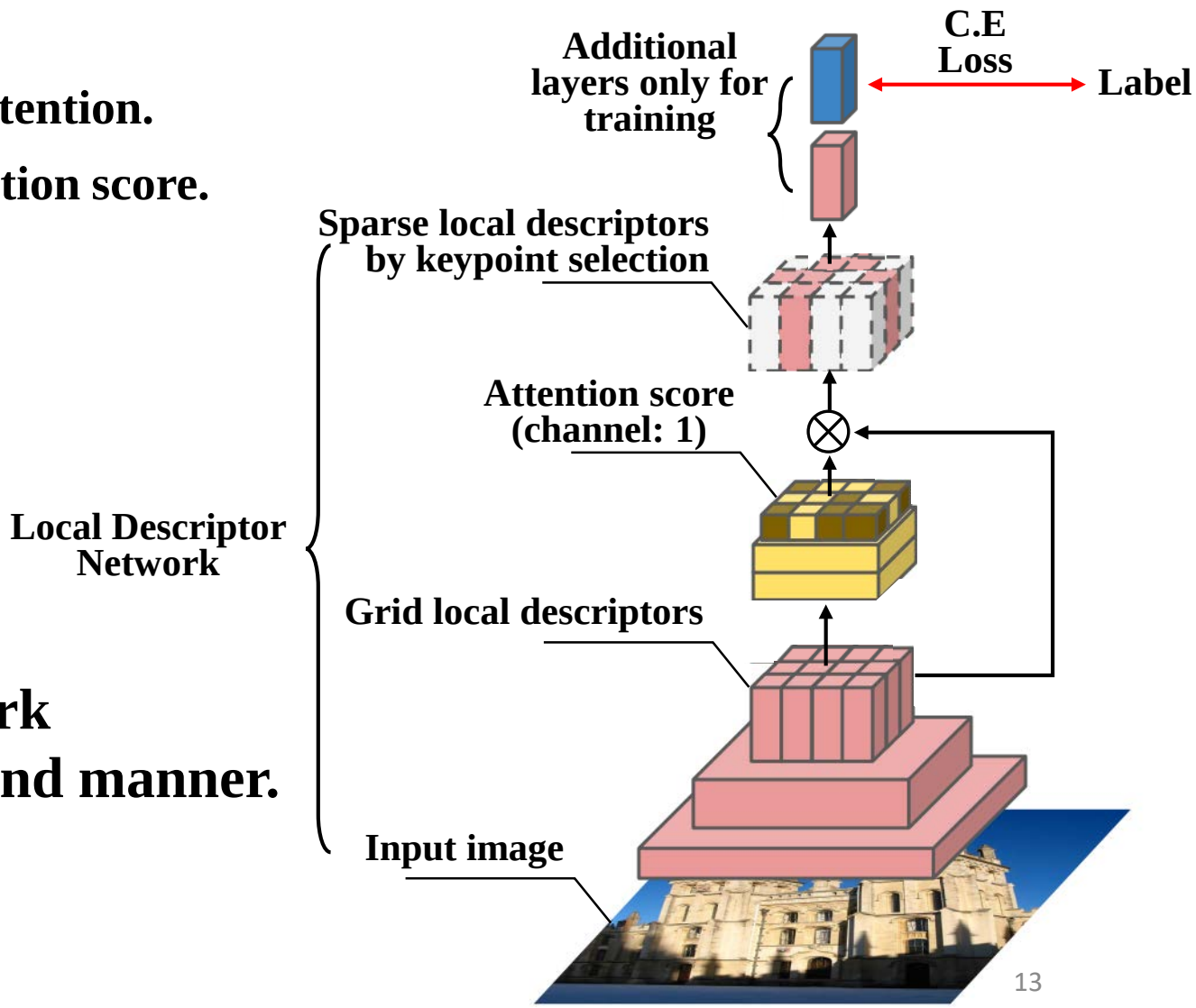
- Backbone: ResNet50 trained on ImageNet
- Fine tune the local descriptor network on Google Landmarks dataset.
  (Image Retrieval dataset)
- Use only cross-entropy loss (loss of classification task)
- However, this method doesn't perform keypoint selection.



Training for Keypoint Selection

- Keypoint selection can be performed by attention.
- Attention module is used to calculate attention score.
- Attention score is close to 0: no keypoint
- Attention score is close to 1: keypoint

∴ We can train the local descriptor network performing keypoint selection by end-to-end manner.



Comparison with classical local descriptor

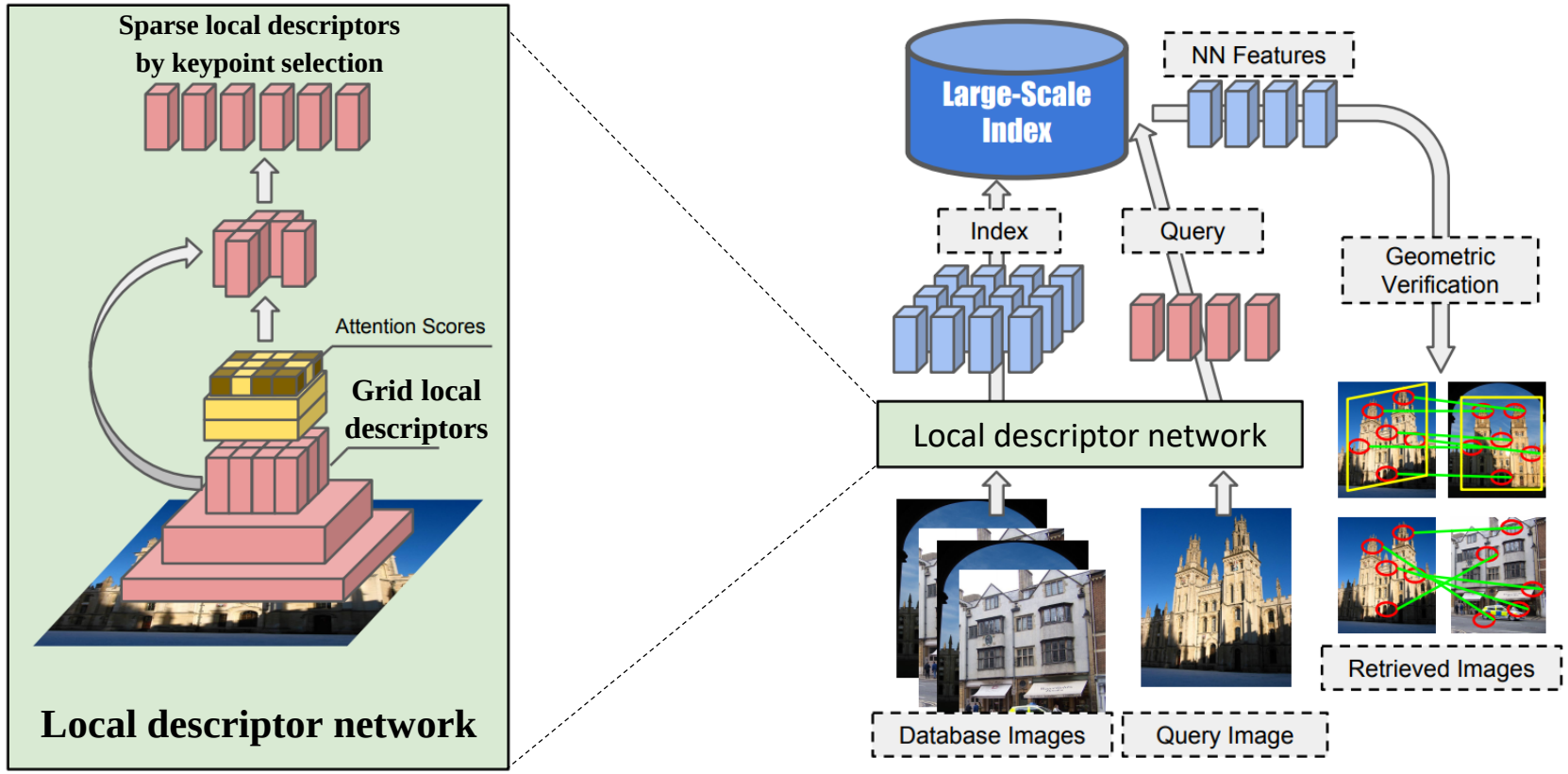
- **Classical Local descriptor**
  - **① keypoint selection**
  - **② local descriptor extraction**

- **Ours**
  - **① local descriptor extraction**
  - **② keypoint selection**

- **The order of process is different, but the results are similar.**

Image retrieval pipeline with ours

- **Overall image retrieval pipeline with our local descriptor network**

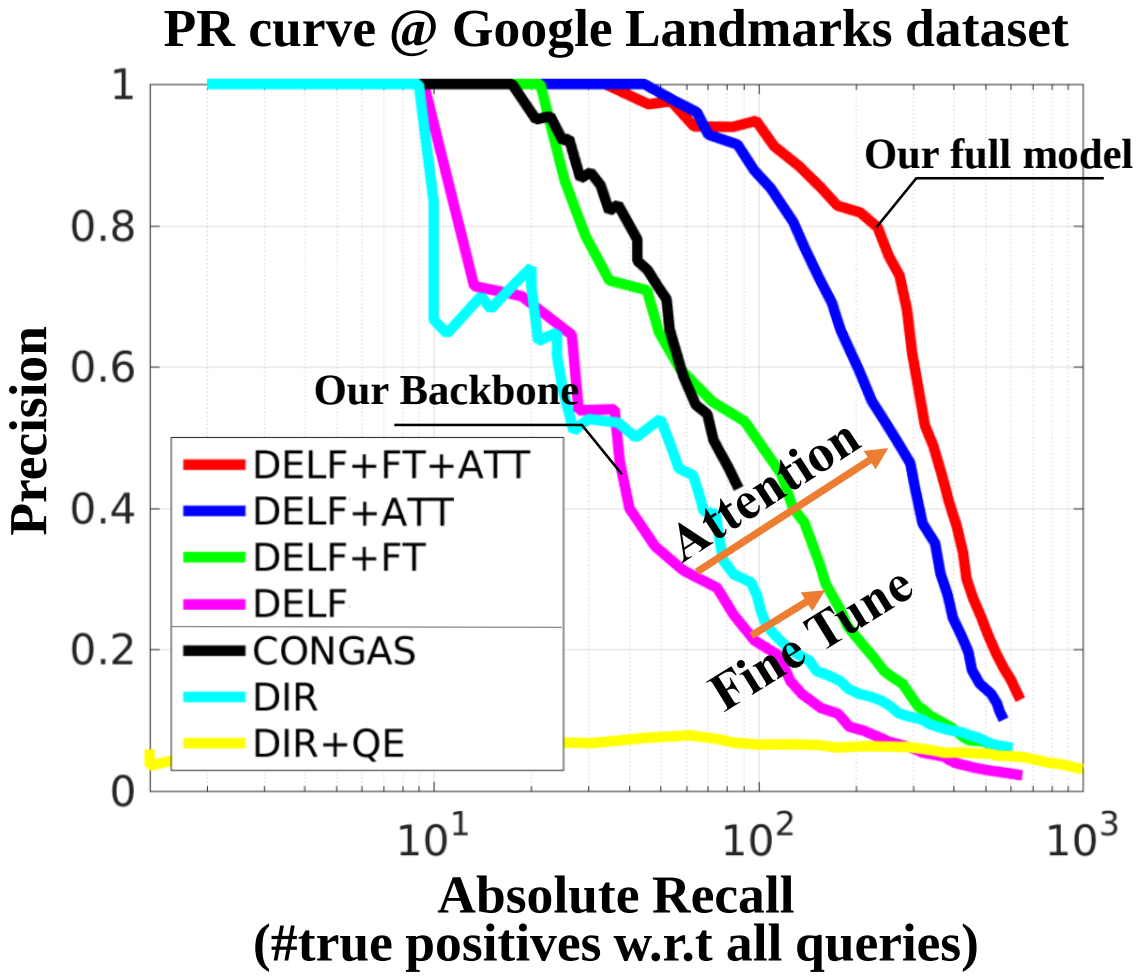


Contents

- **Background & Motivation**
- **Our Approach**
- **Results**
- **Summary**
- **Quiz**

Precision & Recall Result

- DELF: Backbone trained on ImageNet
- FT: Fine Tune with Google Landmarks
- ATT: Use Attention for keypoint selection
- QE: Use Query Expansion



mAP Result

- **DELF: Backbone trained on ImageNet**
- **FT: Fine Tune with Google Landmarks**
- **ATT: Use Attention**
- **QE: Use Query Expansion**
- **DIR: Use global descriptor**

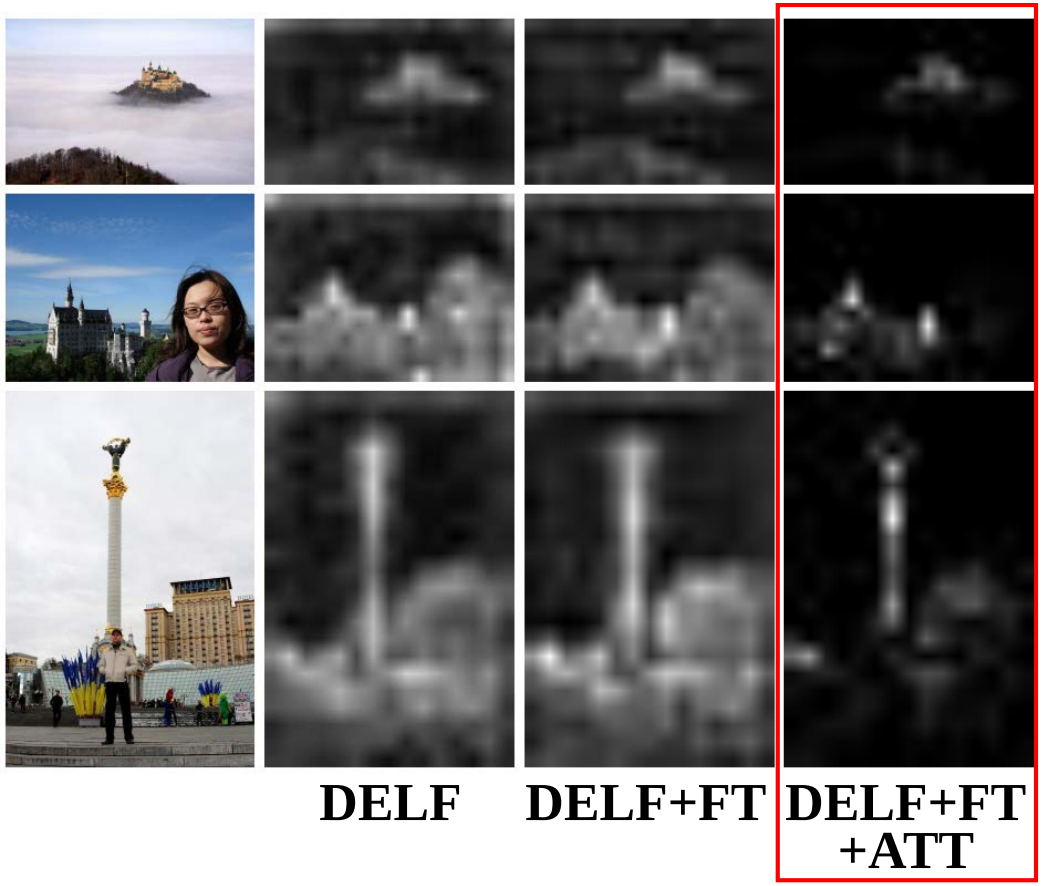
Mean average precision: mAP(%)

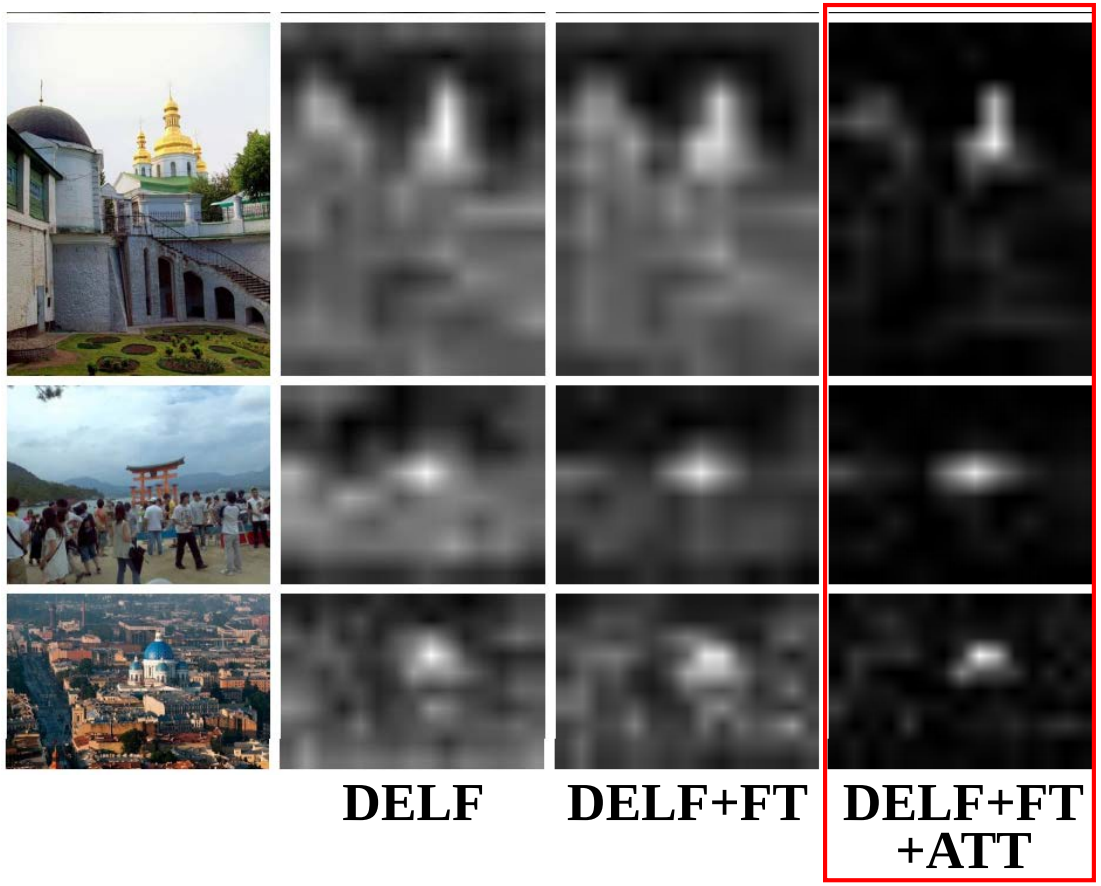
Dataset	Oxf5k	Oxf105k	Par6k	Par106k
DIR [11]	86.1	82.8	94.5	90.6
DIR+QE [11]	87.1	85.2	95.3	91.8
siaMAC [29]	77.1	69.5	83.9	76.3
siaMAC+QE [29]	81.7	76.6	86.2	79.8
CONGAS [8]	70.8	61.1	67.1	56.8
LIFT [40]	54.0	–	53.6	–
DELF+FT+ATT (ours)	83.8	82.6	85.0	81.7
DELF+FT+ATT+DIR+QE (ours)	**90.0**	**88.5**	**95.7**	**92.8**

DELF: Backbone trained on ImageNet
FT: Fine Tune with Google Landmarks
ATT: Use Attention

Keypoint Selection Result(Attention Result)

- Keypoint selection effectively disregards clutter.
- Attention activates on important pixels for image retrieval.





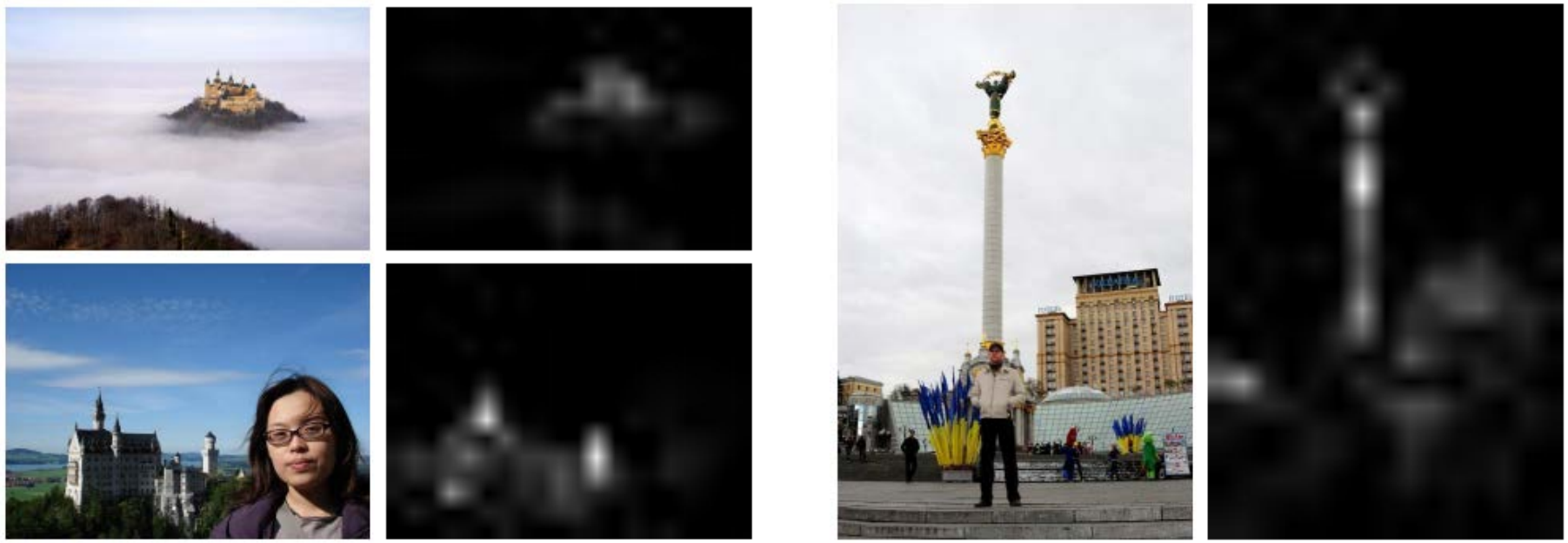
Contents

- **Background & Motivation**
- **Our Approach**
- **Results**
- **Summary**
- **Quiz**

Contributions

- **Keypoint selection using the attention module**
  - **Unnecessary region descriptors are suppressed. (e.g., texture-less region)**
  - **Only sparse local descriptors that are useful for image retrieval are extracted.**

  ∴ Search efficiency ↑, Accuracy ↑



< Result of Keypoint Selection>

Strengths & Weaknesses

- **Strengths**
  - **Proposed method can extract both local descriptors and keypoints via one forward pass.**
  - **Efficient and accurate image retrieval can be performed by keypoint selection using attention.**

- **Weaknesses**
  - **This paper trains the descriptor network using only classification task.**
    (Do not use other metric learning methods)
  - **The image label is required to train the network.**

Contents

- **Background**
- **Related work**
- **Our Approach**
- **Results**
- **Quiz**

Quiz

- **Please submit this google form.**

Link will be posted in the regular zoom meeting session.

THANK YOU