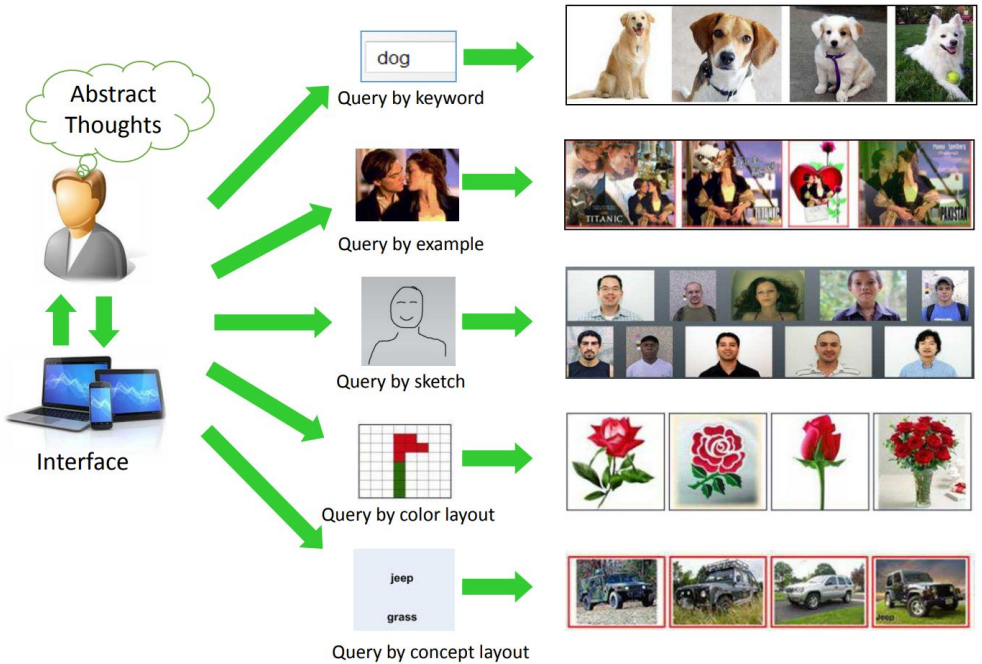
Image Search

Start of CBIR

EAkins, et al..Content-based Image Retrieval, 1999

survey CBIR



Wengang Zhous, et al..Recent Advance in Content-based Image Retrieval: A Literature Survey.2017

Visual Geolocalization

GeoWarp

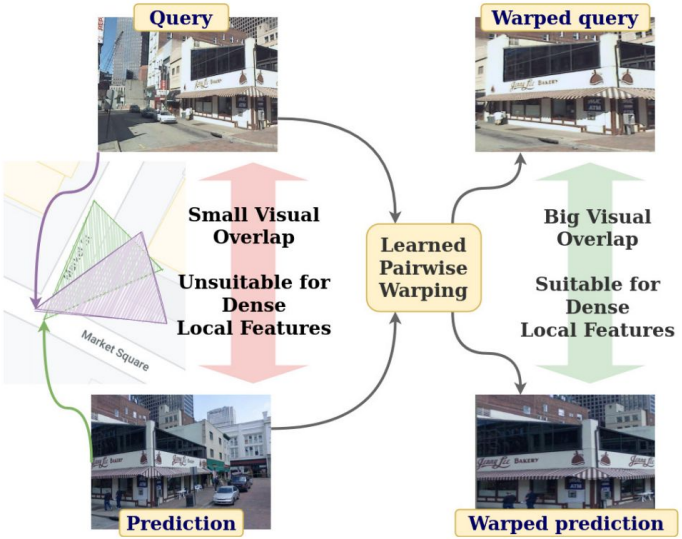
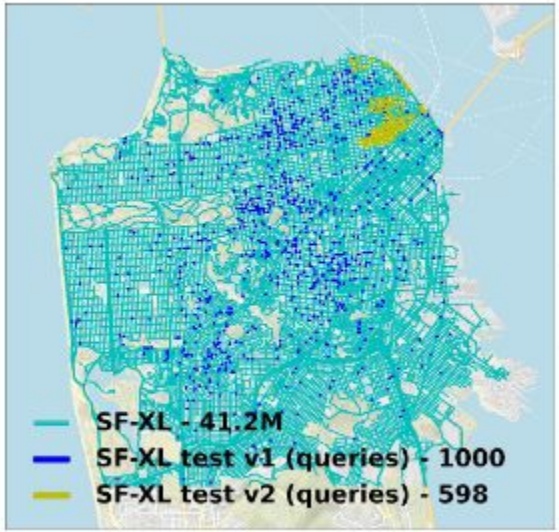
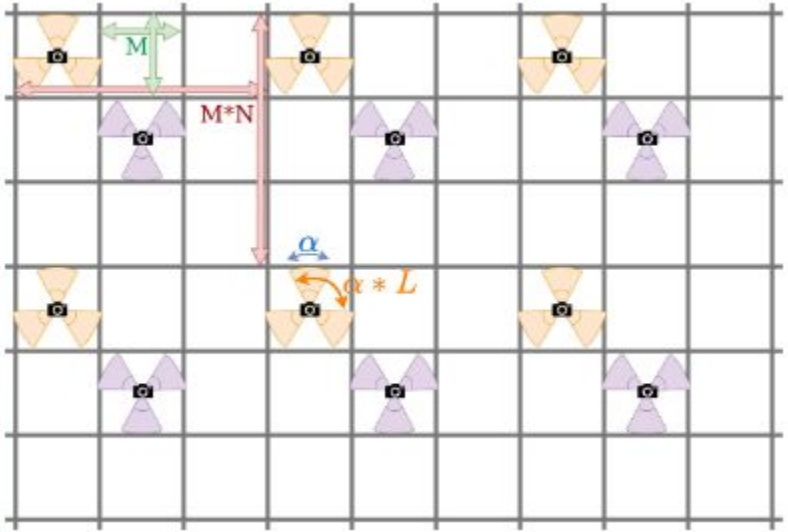


Figure 1. The appearance of two different views of the same place may differ significantly, thus making it hard to match them. Our method warps both images to a closer geometrical space and then computes their similarity using deep dense local features.

Berton, et al..Viewpoint invariant dense matching for visual geolocalization, ICCV 2021

Dataset: SF-XL



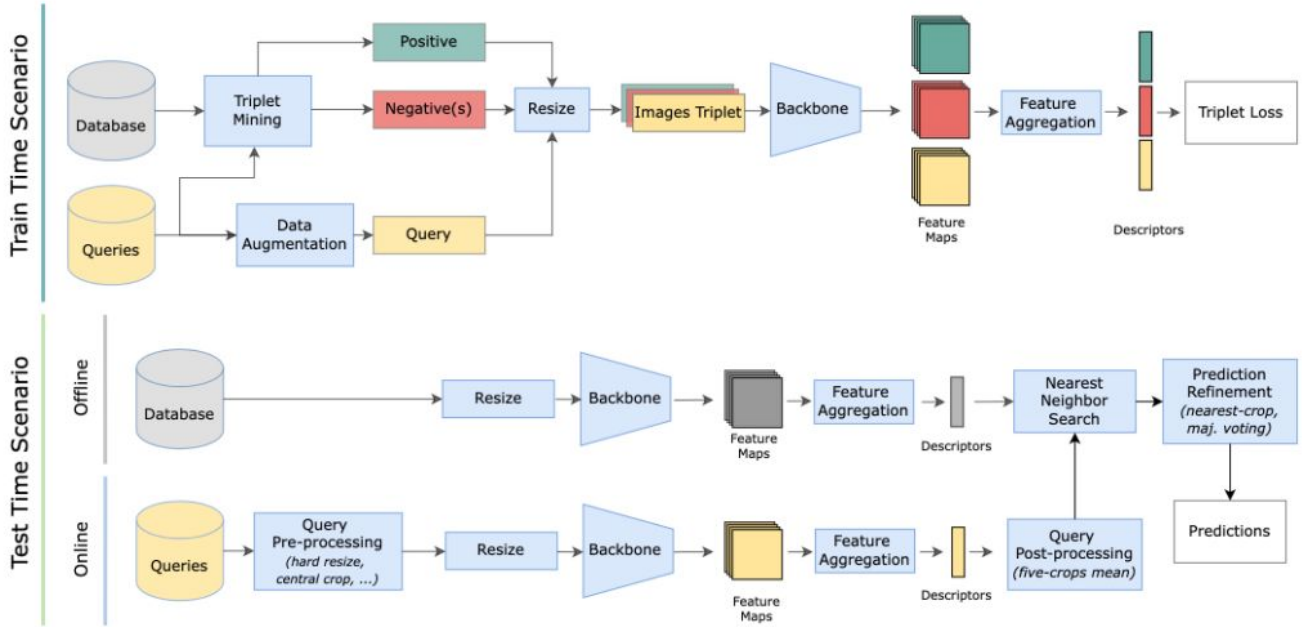


Berton, et al..Rethinking Visual Geo-localization for Large-Scale Application, CVPR 2022

St Lucia

Milford, et al.Mapping a Suburb With a Single Camera Using a Biologically Inspired SLAM System, TRO 2008

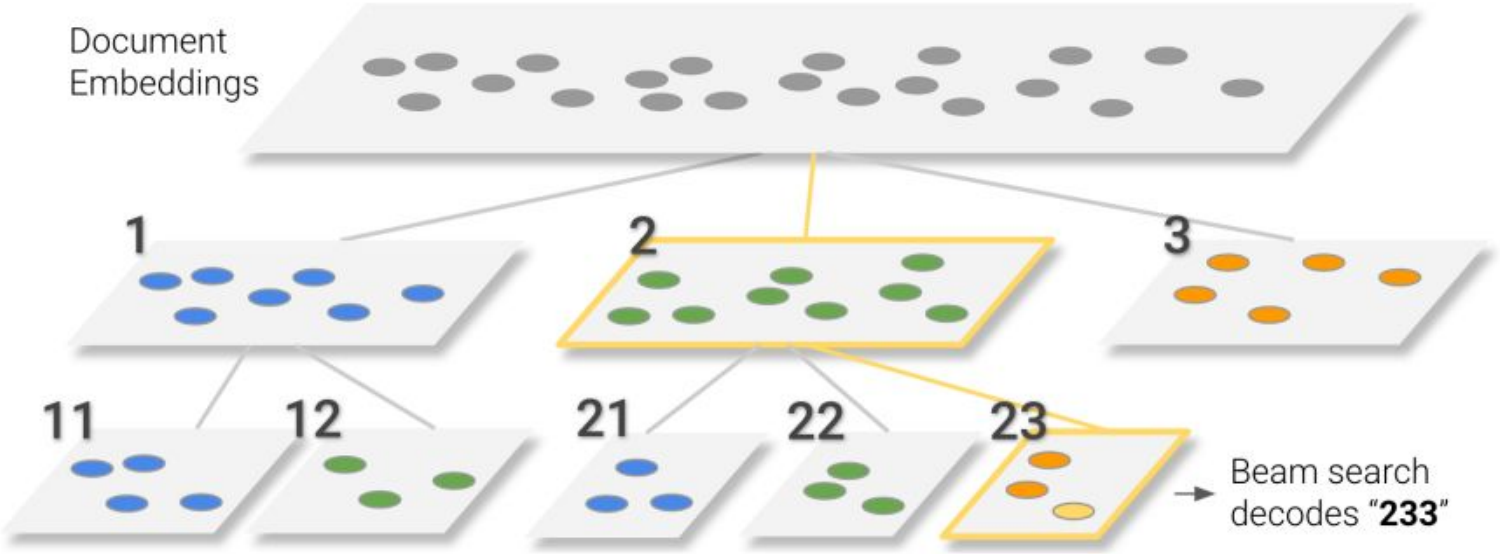
Benchmark



Berton, et al..Deep Visual Geo-localization Benchmark, CVPR 2022

Indexing

Transformer Memory as a Differentiable Search Index

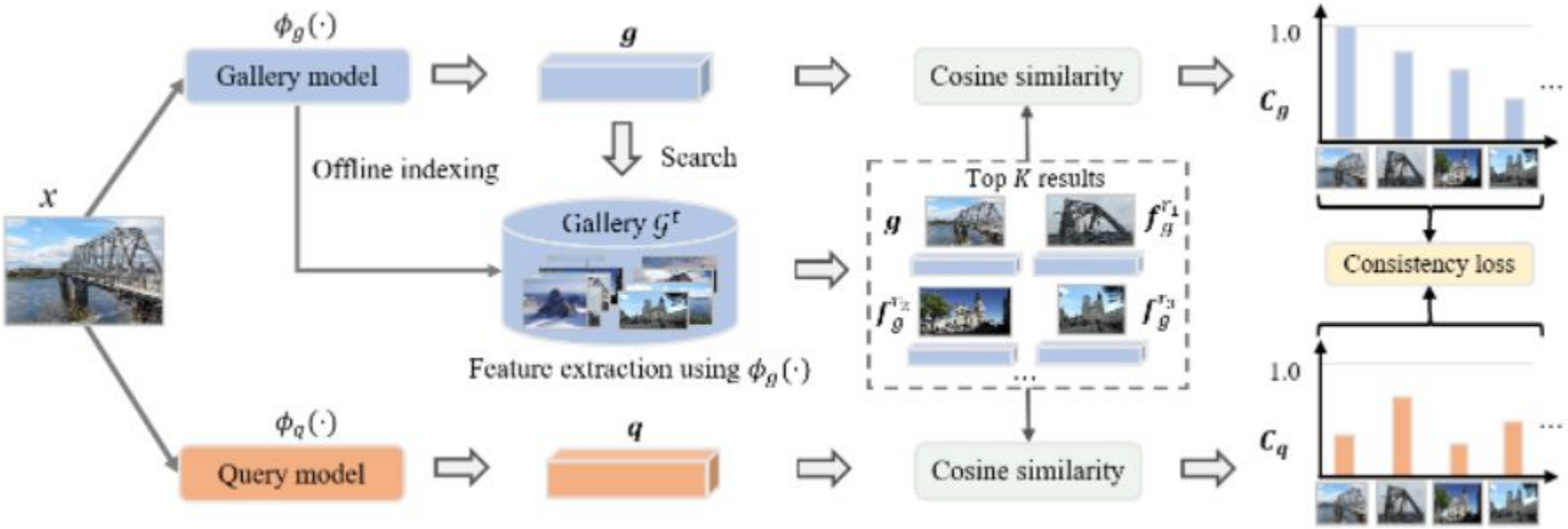


Yi Tay, et al..,Transformer memory as a Differentiable Search Index, 2022

Rerank

1

Contextual Similarity



Hui Wu, et al.., Contextual Similarity Distillation for Asymmetric Image Retrieval, CVPR 2022

Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval

Ondřej Chum[1], James Philbin[1], Josef Sivic[1], Michael Isard[2] and Andrew Zisserman[1]
[1]Visual Geometry Group, Department of Engineering Science, University of Oxford
[2]Microsoft Research, Silicon Valley
{ondra,james,josef,az}@robots.ox.ac.uk misard@microsoft.com

Chum, et al.., Total Recall: Automatic Query Expansion with a Generative Featrue Model for Object Retrieval, ICCV 2007

K-reciprocal nearest neighbors

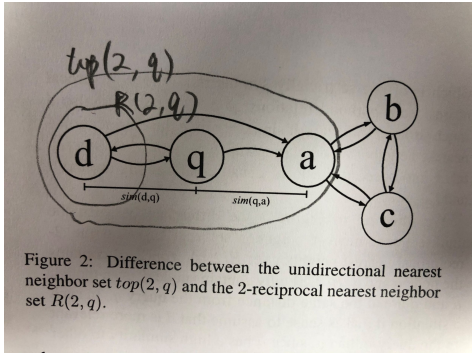
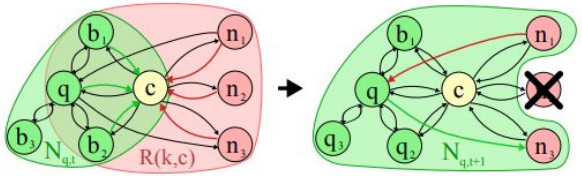


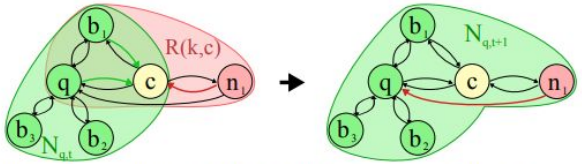
Figure 2: Difference between the unidirectional nearest neighbor set $top(2, q)$ and the 2-reciprocal nearest neighbor set $R(2, q)$.

Make a close set for every query or every db image.

- 互相离得近的是close set, 否则是far set
- Far set中和close set近，和其他图片远的比较重要



(a) Node C's neighborhood is considered, if it contains more than the half of the initial set, *i.e.* $|N_{q,t} \cap \mathrm{R}(k, c)| > \frac{1}{2} |N_{q,t}|$



(b) Node C's neighborhood is considered, if it adds less unknown nodes than known, *i.e.* $|N_{q,t} \cap \mathrm{R}(k, c)| > |\mathrm{R}(k, c) \setminus N_{q,t}|$

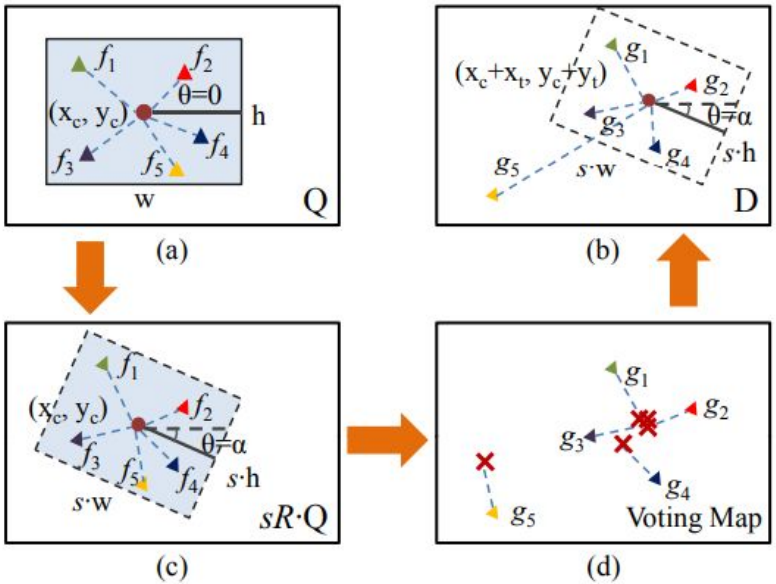
Algorithm 1: Expansion step

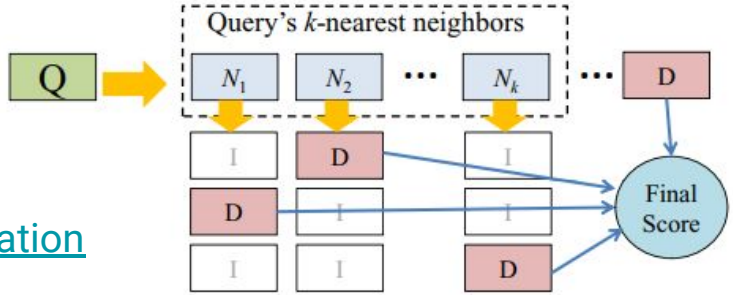
```
for t ← 0 to 2 do
    N_{q,t+1} ← N_{q,t};
    foreach n ∈ N_{q,t} do
        if |N_{q,t} ∩ R(k,n)| > 1/2 |N_{q,t}| then
            N_{q,t+1} ← R(k,n) ∪ N_{q,t+1};
        if |N_{q,t} ∩ R(k,n)| > |R(k,n) \ N_{q,t}| then
            N_{q,t+1} ← R(k,n) ∪ N_{q,t+1};
```

The first condition allows only nodes which are connected to at least half of the *close set* to bring in their neighbors. This high connectivity ensures that added nodes are very likely to be relevant to the query. The second condition relaxes this restriction slightly by allowing weakly connected nodes to bring in their neighbors if the amount of new neighbors is smaller than the amount of connections already made to the *close set*. Nodes added to $N_{q,t+1}$ are sorted according to $sim(q, d)$ and inserted in this order into

Qin, et al.., Hellor neighbor: accurate object retrieval with k-reciprocal nearest neighbors, CVPR 2011

Spatially Constrained Similarity Measure



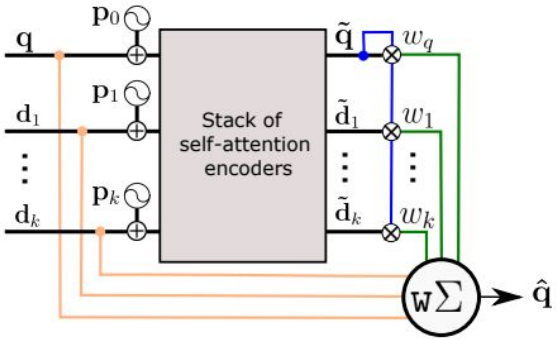


[US8880563B2 Image search by query object segmentation]

[Shen, et al.., Spatially-Constrained Similarity Measure for Large-Scale Object Retrieval, TPAMI 2014]

[Shen, et al.., Object retrieval and localization with spatially-constrained similarity measure and k-NN re-ranking, CVPR 2012 (210)]

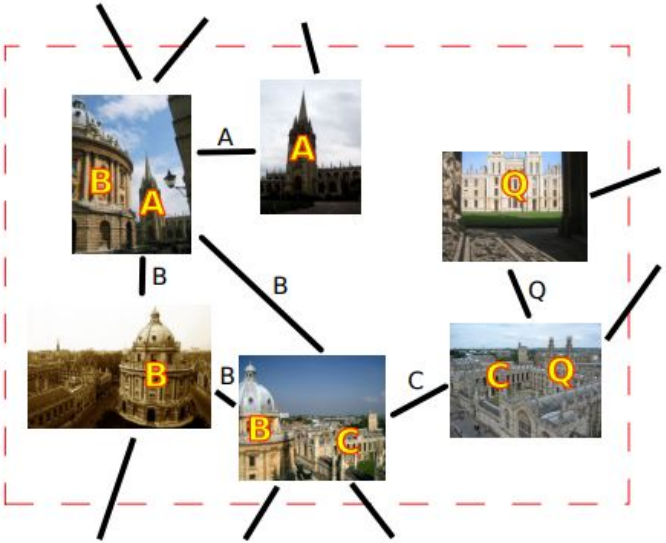
Attention-Based Query Expansion



Method	$\theta(\mathbf{d}_i \mid \mathbf{q}, \{\mathbf{d}\}^k, \{\mathbf{d}\}^-, i) = w_i \mathbf{d}_i$
[7] **AQE:** Average QE	$w_i = 1$
[13] **AQEwD:** AQE with decay	$w_i = (k - i)/k$
[3] **DQE:** Discriminative QE	$\mathbf{w}$ is the dual-form solution of an SVM optimization problem using $\{\mathbf{d}\}^k$ as positives and $\{\mathbf{d}\}^-$ as negatives
[33] **αQE:** α-weighted QE	$w_i = \text{sim}(\mathbf{q}, \mathbf{d}_i)^\alpha$, with α being a hyperparameter.

Gordo, et al.., Attention-Based Query Expansion Learning, ECCV 2020

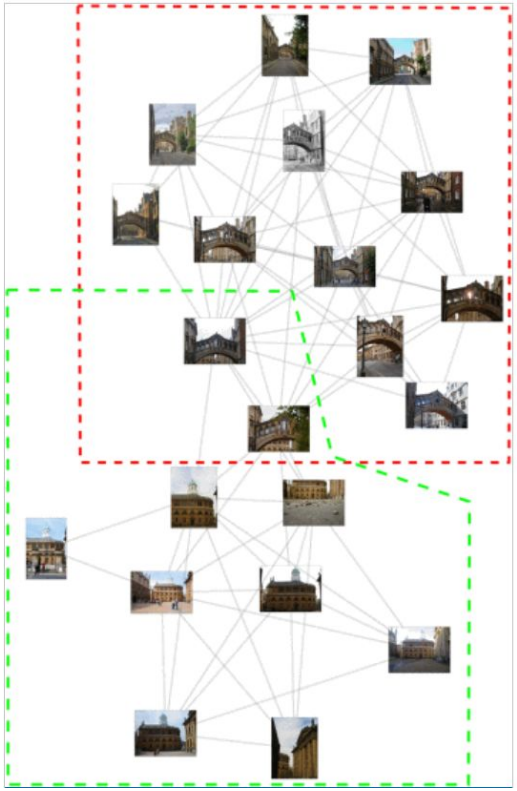
Database Augmentation (before CNN)



root SIFT
QE+SVM
DBA

Arandjelovic, et al.., Three things everyone should know to improve object retrieval, CVPR 2012

Matching graph



philbin, et al.., Object Mining Using a Matching Graph on Very Large Image Collections, 2008 Sixth Indian Conference on Computer Vision, Graphics \& Image Processing

QE

Local features: [Aggregate13Tolias],[Three12Arandjelovic]

Geometric verification

3

Improve Spatial verification

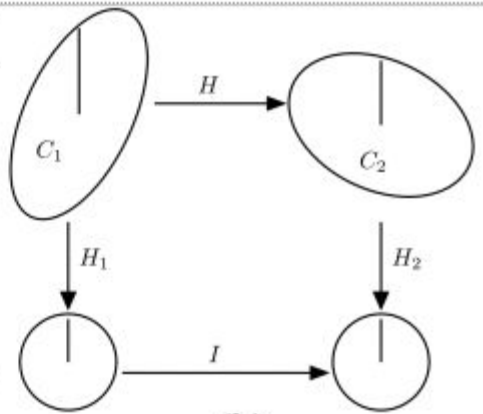
Raguram, et al.., Improved Geometric Verification for Large Scale Landmark Image Collection, BMVC 2012

Spatial verification

Table 6. (a) The three affine sub-groups compared in the spatial re-ranking. (b) Computing H as $H_2^{-1} H_1$, preserving "upness" for the 5 dof case.

Transformation	dof	Matrix
translation + isotropic scale	3	$\begin{bmatrix} a & 0 & t_x \\ 0 & a & t_y \end{bmatrix}$
translation + anisotropic scale	4	$\begin{bmatrix} a & 0 & t_x \\ 0 & b & t_y \end{bmatrix}$
translation + vertical shear	5	$\begin{bmatrix} a & 0 & t_x \\ b & c & t_y \end{bmatrix}$

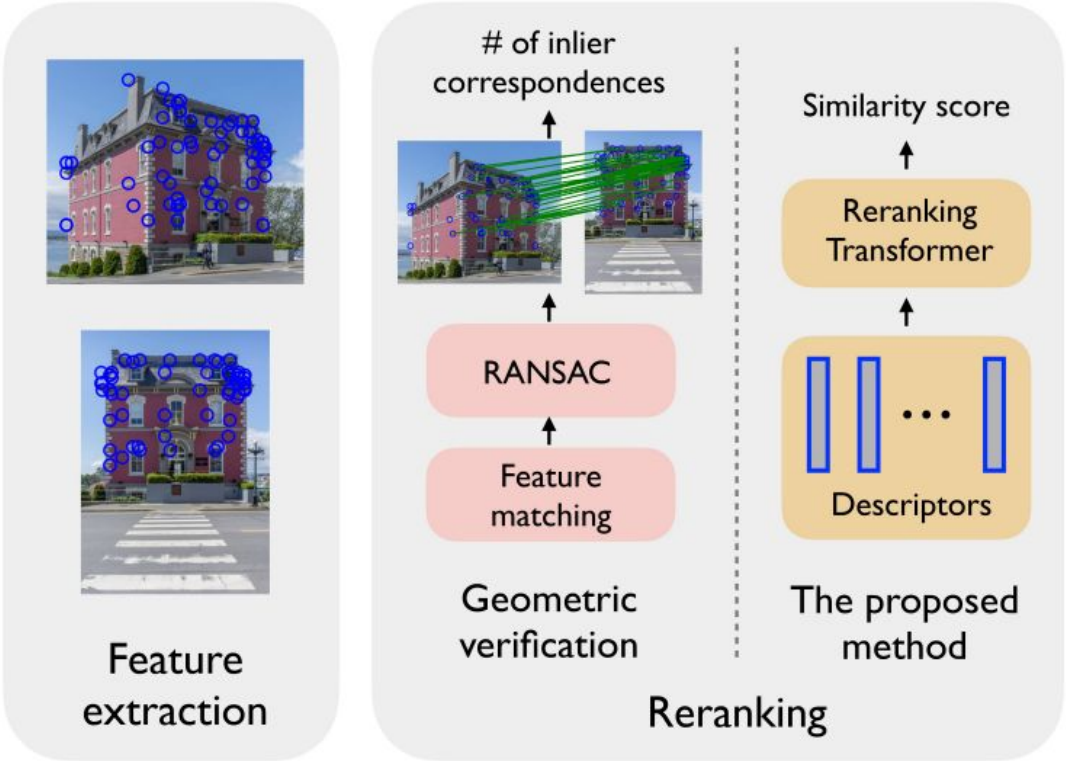
(a)



(b)

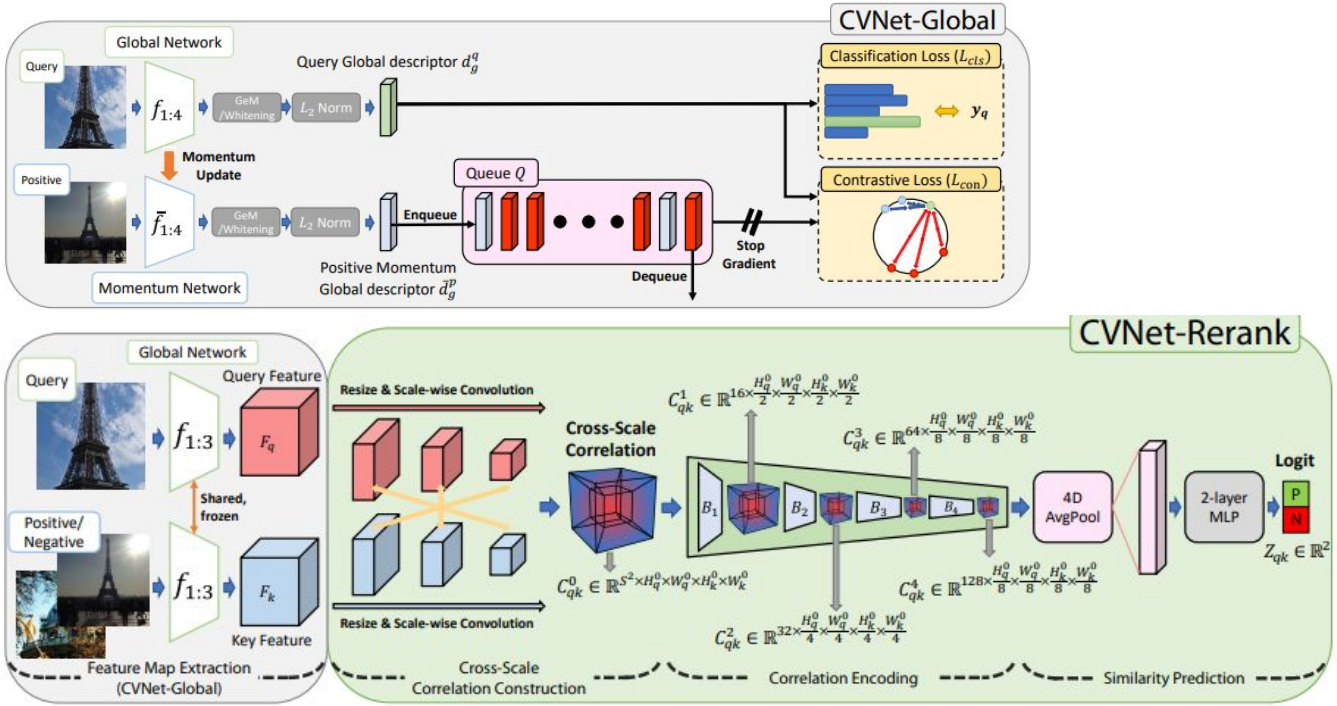
Philbin, et al.., Object retrieval with large vocabularies and fast spatial matching, CVPR 2007

Reranking transformers



Tan, et al.., Instance-level Image Retrieval using Reranking Transformers, ICCV 2021

Correlation Verification



Lee, et al.., Correlation Verification for Image Retrieval, CVPR 2022

Diffusion

3

Diffusion in semi-supervised learning

$$F(t+1) = \alpha S F(t) + (1-\alpha) Y$$

$$F^* = \lim_{t\to\infty} F(t) = (1-\alpha)(I-\alpha S)^{-1} Y,$$

贡献是提出使用 $S = D^{-1/2} W D^{-1/2}$

Rigorous proof

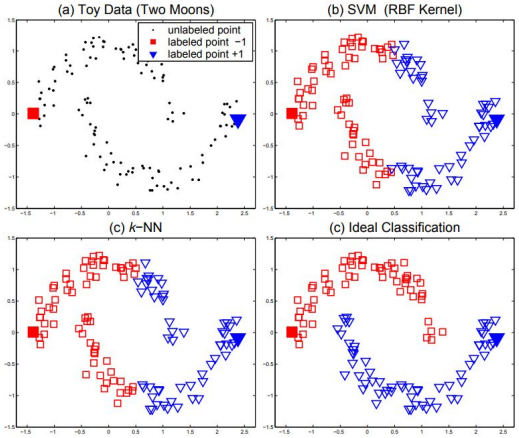
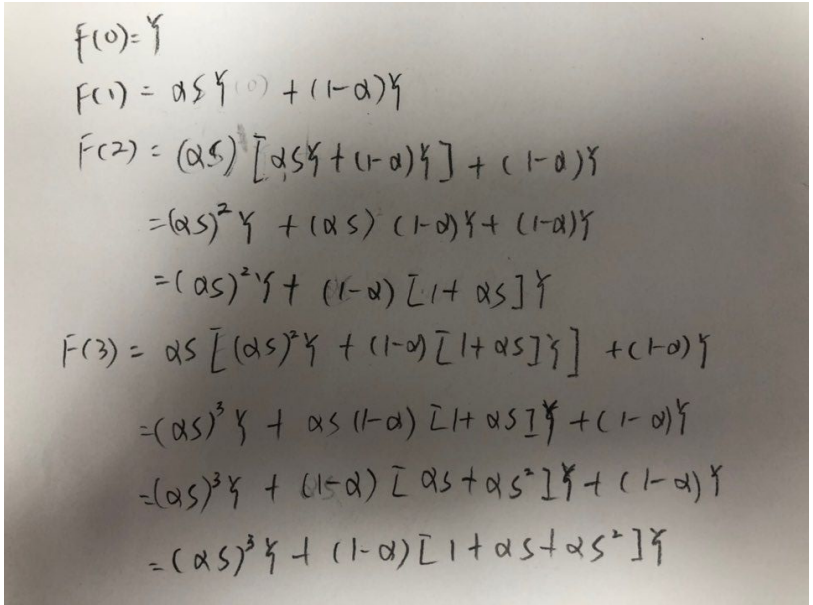


Figure 1: Classification on the two moons pattern. (a) toy data set with two labeled points; (b) classifying result given by the SVM with a RBF kernel; (c) k-NN with $k = 1$; (d) ideal classification that we hope to obtain.



Zhou, et al.., Learning with Local and Global Consistency, NIPS 2004 (3997)

Diffusion in Image/Text Ranking

贡献是提出使用 $S = D^{-1/2} W D^{-1/2}$

$$F(t+1) = \alpha S F(t) + (1-\alpha) Y$$

$$F^* = \lim_{t \to \infty} F(t) = (1-\alpha)(I - \alpha S)^{-1} Y,$$

Easy proof

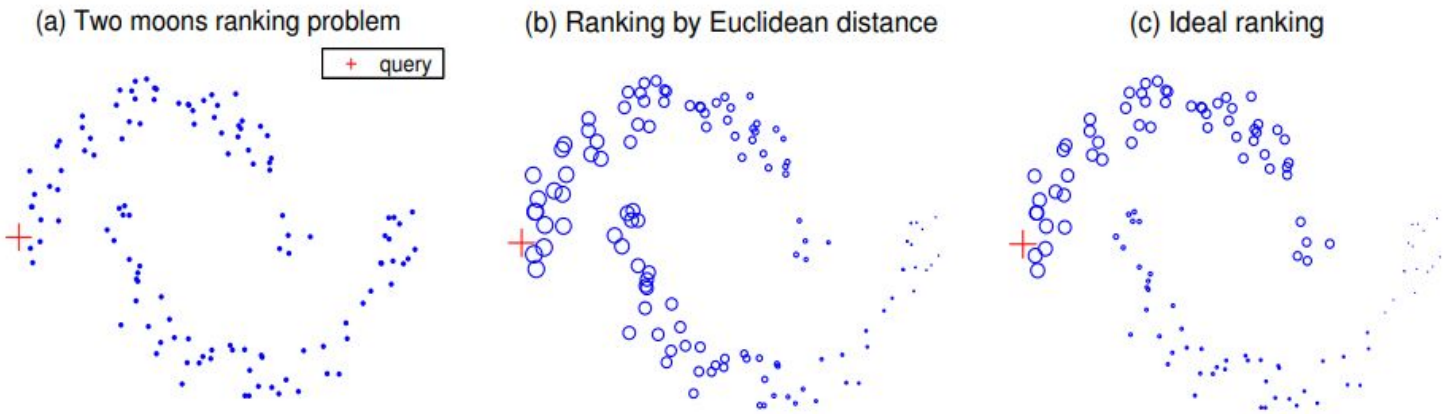


Figure 1: Ranking on the two moons pattern. The marker sizes are proportional to the ranking in the last two figures. (a) toy data set with a single query; (b) ranking by the Euclidean distances; (c) ideal ranking result we hope to obtain.

Zhou, et al.,.Ranking on Data Manifold, NIPS 2004 (812)

Diffusion and Personalized PageRank

PageRank

$$P = (1-\epsilon)U + \epsilon D^{-1}W,$$

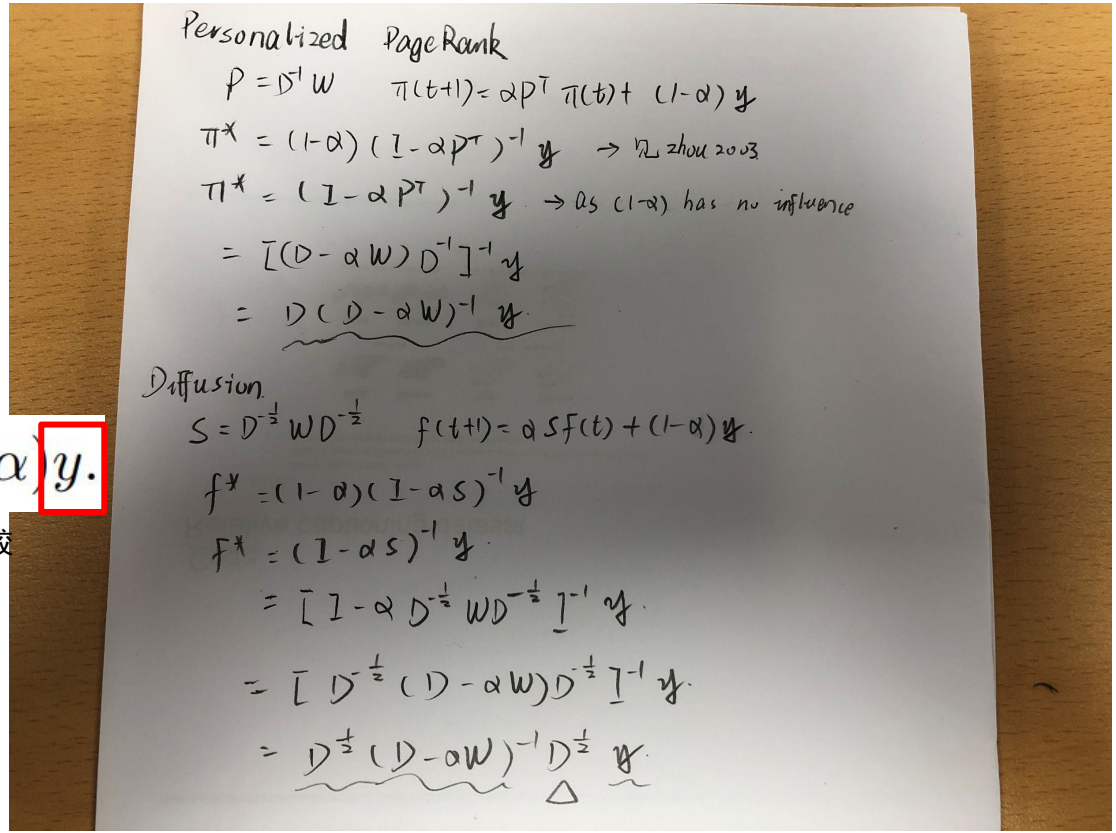
Personalized PageRank

$$P = D^{-1}W.$$

$$\pi(t+1) = \alpha P^T \pi(t) + (1-\alpha)y.$$

区别是U是uniformed, y可以给感兴趣的网页赋予较大的值

Note that $P = D^{-1}W.$ $P = (1-\epsilon)U + \epsilon D^{-1}W,$ have same stationary distribution



Zhou, et al..,Ranking on Data Manifold, NIPS 2004 (812)

Page, et al..,The PageRank CitationRanking: Bringing Order to the Web, Stanford InfoLab 1999 (14107)

Graph Transduction

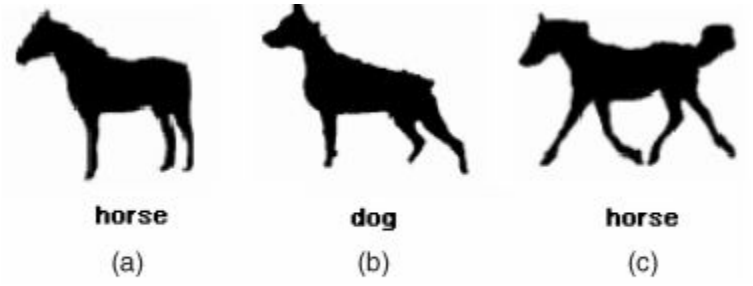


Fig. 1. Existing shape similarity methods incorrectly rank shape (b) as more similar to (a) than (c).

the original similarities $w_{i,j} = sim(x_i, x_j)$. Our intuition is that the new similarity $f(x_i) = s(x_1, x_i)$ will be large iff all points x_j that are very similar to x_i (large $sim(x_i, x_j)$) are also very similar to query x_1 (large $sim(x_1, x_j)$). Note that

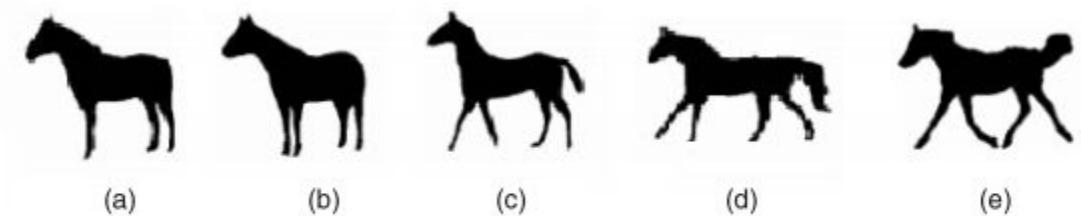


Fig. 2. A key idea of the proposed distance learning is to replace the original shape distance between (a) and (e) with a distance induced by geodesic paths in the manifold of known shapes. One such path is (a)-(e) in this figure.

Bai, et al..,Learning Context-Sensitive Shape Similarity by Graph Transduction, PAMI 2009 (296)

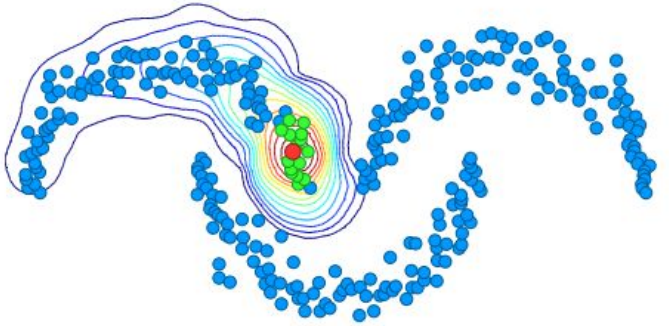
Summary of diffusion processes

This paper uses clear math terms, so it is helpful to search and review the mathematical knowledge.

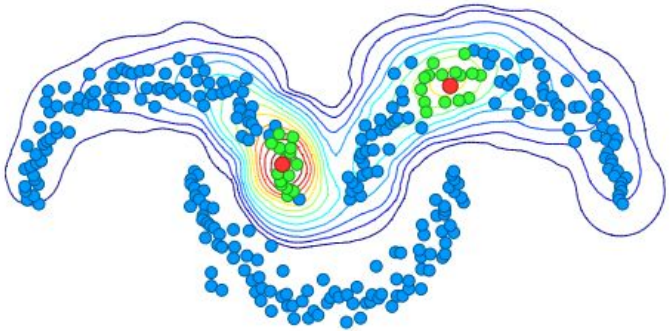
Method	Abbr.	Initialization $\mathbf{W}^0$	Transition $\mathbf{T}$	Diffusion
Global PageRank [17]	GPR	$\mathbf{u}$	$\mathbf{P}$	$\mathbf{f}_{t+1} = \mathbf{f}_t\,\mathbf{T}$
Personalized PageRank [17]	PPR	$\mathbf{u}$	$\mathbf{P}$	$\mathbf{f}_{t+1} = \alpha\,\mathbf{f}_t\,\mathbf{T} + (1-\alpha)\,\mathbf{y}$
Ranking on Manifolds [23]	ROM	$\mathbf{u}$	$\mathbf{P}_{\text{NC}}$	$\mathbf{f}_{t+1} = \alpha\,\mathbf{f}_t\,\mathbf{T} + (1-\alpha)\,\mathbf{y}$
Label Propagation [24]	LP	$\mathbf{y}$	$\mathbf{P}$	$\mathbf{f}_{t+1} = \mathbf{f}_t\,\mathbf{T}$ and $f\,(i) = 1$
Graph Transduction [2]	GT	$\mathbf{y}$	$\mathbf{P}$	$\mathbf{f}_{t+1} = \mathbf{f}_t\,\mathbf{T}$ and $f\,(i) = 1$
Locally Constrained DP [21]	LCDP	$\mathbf{A}$	$\mathbf{P}_{\text{kNN}}$	$\mathbf{W}_{t+1} = \mathbf{T}\,\mathbf{W}_t\,\mathbf{T}^T$
Tensor Graph Diffusion [22]	TGD	$\mathbf{A}$	$\mathbf{P}_{\text{DS}}$	$\mathbf{W}_{t+1} = \mathbf{T}\,\mathbf{W}_t\,\mathbf{T}^T + \mathbf{I}$
Shortest Path Propagation [20]	SPP	$\mathbf{y}$	$\mathbf{P}_{\text{SP}}$	$\mathbf{f}_{t+1} = \mathbf{f}_t\,\mathbf{T}$
Self Smoothing Operator [8]	SSO	$\mathbf{A}$	$\mathbf{P}$	$\mathbf{W}_{t+1} = \mathbf{W}_t\,\mathbf{T}$
Self Diffusion [19]	SD	$\mathbf{A}$	$\mathbf{P}$	$\mathbf{W}_{t+1} = \mathbf{W}_t\,\mathbf{T} + \mathbf{I}$
Replicator Dynamics [18]	RD	$\mathbf{u}$	$\mathbf{A}$	$\mathbf{f}_{t+1} = \mathbf{f}_t \odot \mathbf{T}\,\mathbf{f}_t$ and $\mathbf{f}_{t+1} = \mathbf{f}_{t+1} / \lvert\mathbf{f}_{t+1}\rvert$
Power Iteration Clustering [11]	PIC	$\mathbf{s}$	$\mathbf{P}$	$\mathbf{f}_{t+1} = \mathbf{T}\,\mathbf{f}_t$ and $\mathbf{f}_{t+1} = \mathbf{f}_{t+1} / \lvert\mathbf{f}_{t+1}\rvert$
Authority Shift Clustering [3]	ASC	$\mathbf{P}_{\text{PPR}}$	$\mathbf{P}_{\text{PPR}}$	$\mathbf{W}_{t+1} = \mathbf{W}_t\,\mathbf{T}$

Donoser, et al..,Diffusion Processes for Retrieval Revisited, CVPR 2013 (174)

Efficient; region manifolds; small objects



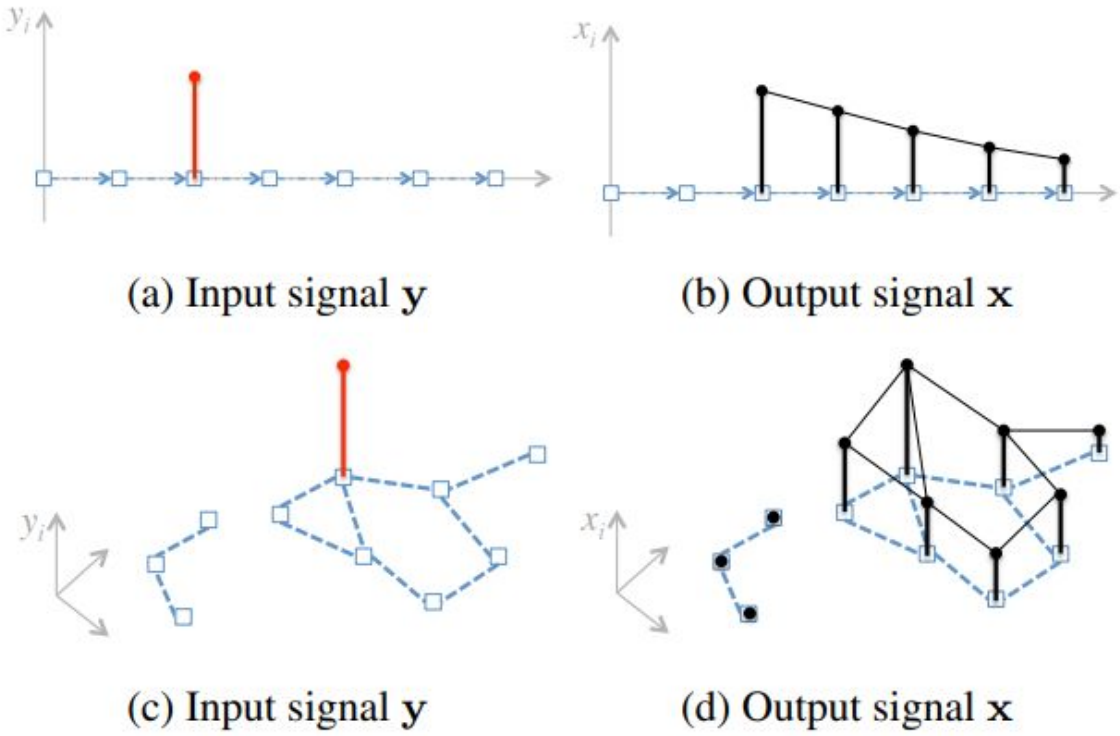
(a) single query



(b) multiple queries

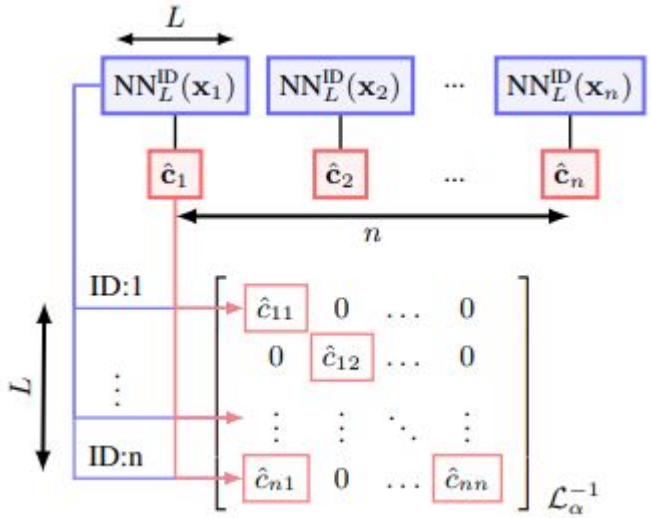
Iscen, et al...Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations, CVPR 2017 (104)

Fast spectral ranking



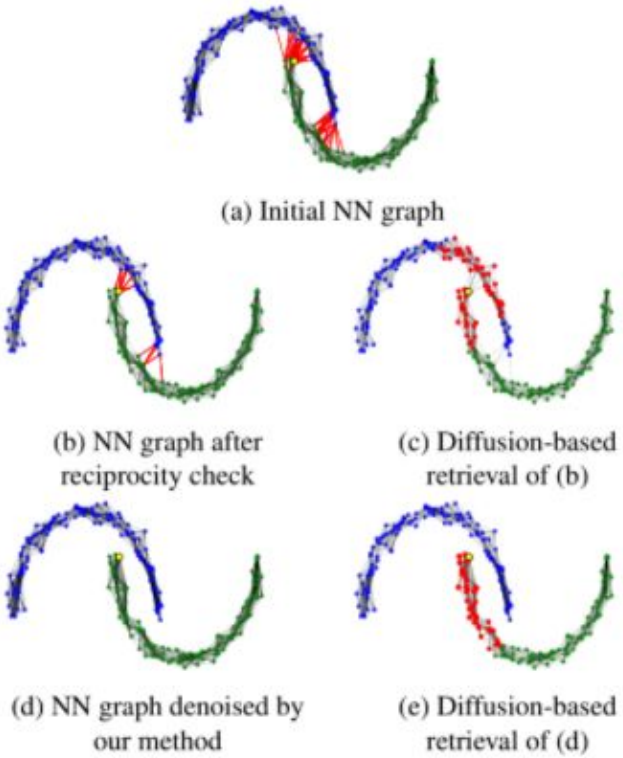
Iscen, et al...Fast Spectral Ranking for Similarity Search, CVPR 2018

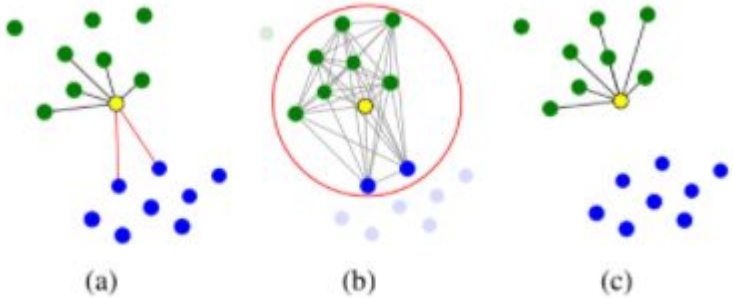
Offline diffusion



Yang, et al...Efficient Image Retrieval via Decoupling Diffusion into Online and Offline Processing, AAAI 2019

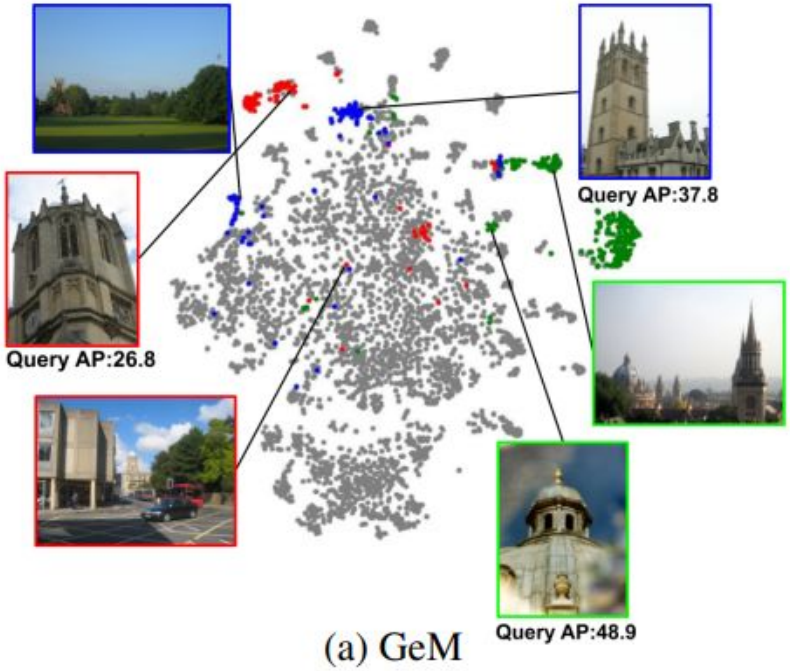
Noisy edges





Kim, et al..,Denoising Nearest Neighbor Graph via Continuous CRF for Image Retrieval, submit to WACV 2020

Guided Similarity Separation



Liu, et al..,Guided Similarity Separation for Image Retrieval NIPS 2019

Hypergraph diffusion

$$X(t) = X(0)P^t = X(0)\,AB \ldots AB = X(0)AQ^{t-1}B$$

$$P = D_v^{-1} W D_e^{-1} W^T = AB.$$

: *vertex-edge transition matrix* $A = D_v^{-1} W$

edge-vertex transition matrix $B = D_e^{-1} W^T$.

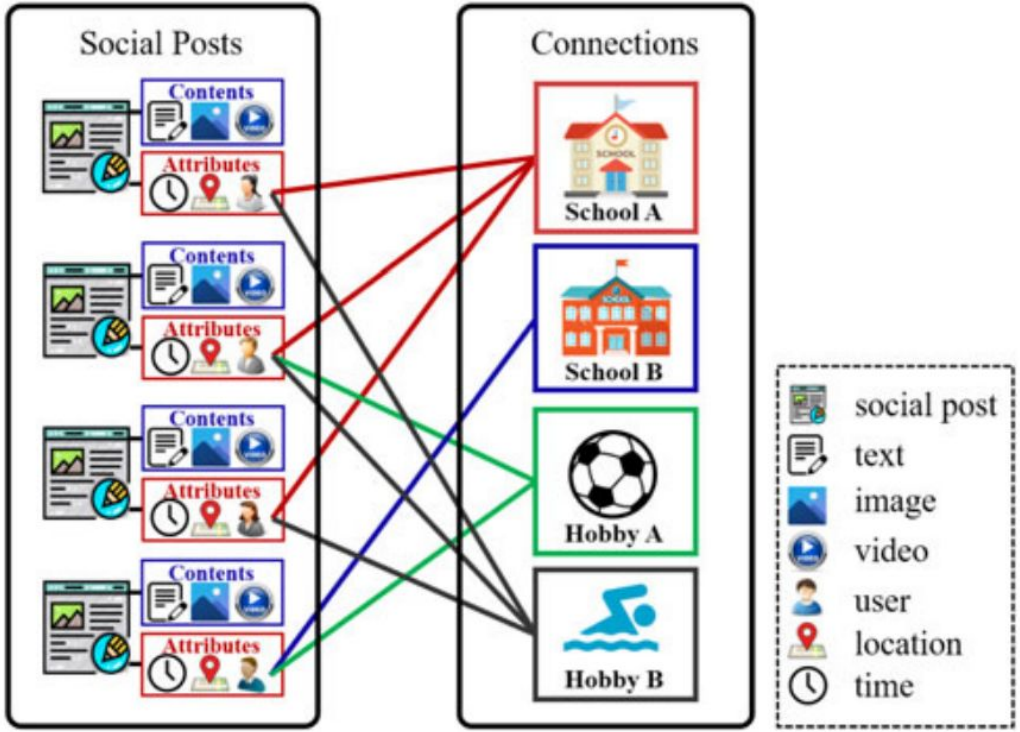
Avin, et al..,Radio cover time in hyper-graphs Ad Hoc Networks 2014

Louis, et al..,Hypergraph Markov Operators, Eigenvalues and Approximation Algorithms STOC 2014

Chan, et al..,Generalizing the Hypergraph Laplacian via a Diffusion Process with Mediators TCS 2020

Hayashi, et al..,Hypergraph Random Walks, Laplacians, and Clustering CIKM 2020

Hypergraph NN

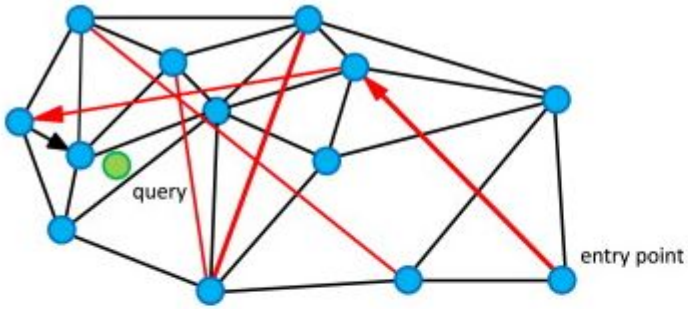


Yifan Feng, et al..,HGNN+: General Hypergraph Neural Networks TPAMI 2022

Retrieval Method

2

NSW



Download : Download full-size image

Fig. 1. Graph representation of the structure. Circles (vertices) are the data in metric space, black edges are the approximation of the Delaunay graph, and red edges are long range links for logarithmic scaling. Arrows show a sample path of the greedy algorithm from the entry point to the query (shown green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Malkov, et al.., Approximate nearest neighbor algorithm based on navigable small world graphs, Information Systems 2014

Hierarchical NSW

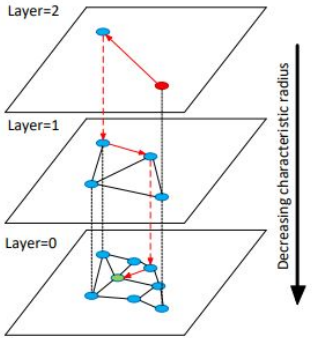


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

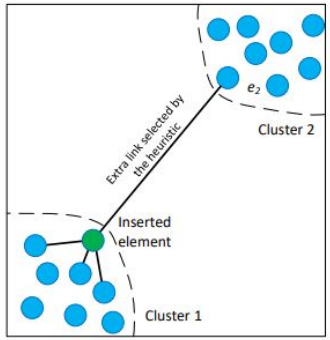
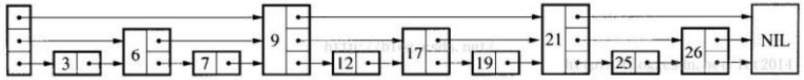
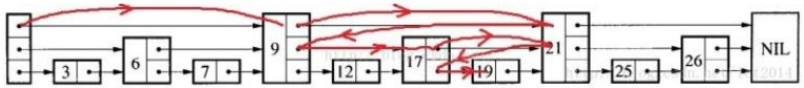


Fig. 2. Illustration of the heuristic used to select the graph neighbors for two isolated clusters. A new element is inserted on the boundary of Cluster 1. All of the closest neighbors of the element belong to the Cluster 1, thus missing the edges of Delaunay graph between the clusters. The heuristic, however, selects element e_2 from Cluster 2, thus, maintaining the global connectivity in case the inserted element is the closest to e_2 compared to any other element from Cluster 1.



查找

查找示意图如下:



比如我们要查找key为19的结点，那么我们不需要逐个遍历，而是按照如下步骤:

- 从header出发，从高到低的level进行查找，先索引到9这个结点，发现9 < 19,继续查找(然后在level==2这层)，查找到21这个节点，由于21 > 19, 所以结点不往前走，而是level由2降低到1
- 然后索引到17这个节点，由于17 < 19, 所以继续往后，索引到21这个结点，发现21>19, 所以level由1降低到0
- 在结点17上，level==0索引到19,查找完毕。
- 如果在level==0这层没有查找到，那么说明不存在key为19的节点，查找失败

Malkov, et al.., Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small Wold graphs, TPAMI 2018

Diffusion on Region Manifolds

- Contributions
  -

Iscen, et al.., Efficient Diffusion on Region Manifolds: Recovering Small Objects with Compact CNN Represenations, CVPR 2017

Aggregation

3

ASMK (aggregated selective match kernels)

$$\mathrm{K}(\mathcal{X},\mathcal{Y}) = \gamma(\mathcal{X})\,\gamma(\mathcal{Y}) \sum_{c\in\mathcal{C}} w_c\, \mathrm{M}\left(\mathcal{X}_c,\mathcal{Y}_c\right)$$

BoW $$\mathrm{M}(\mathcal{X}_c,\mathcal{Y}_c) = \#\mathcal{X}_c \times \#\mathcal{Y}_c = \sum_{x\in\mathcal{X}_c}\sum_{y\in\mathcal{Y}_c} 1,$$

BE $$\mathrm{M}\left(\mathcal{X}_c,\mathcal{Y}_c\right) = \sum_{x\in\mathcal{X}_c}\sum_{y\in\mathcal{Y}_c} w\left(h\left(b_x,b_y\right)\right)$$

VLAD $$\begin{aligned}\mathrm{M}\left(\mathcal{X}_c,\mathcal{Y}_c\right) &= V(\mathcal{X}_c)^\top V(\mathcal{Y}_c)\\ &= \sum_{x\in\mathcal{X}_c}\sum_{y\in\mathcal{Y}_c} r(x)^\top r(y).\end{aligned}$$

Non-aggregated kernel

$$\mathrm{M_N}(\mathcal{X}_c,\mathcal{Y}_c) = \sum_{x\in\mathcal{X}_c}\sum_{y\in\mathcal{Y}_c} \sigma\left(\phi(x)^\top\phi(y)\right).$$

$$\mathrm{SMK}(\mathcal{X}_c,\mathcal{Y}_c) = \sum_{x\in\mathcal{X}_c}\sum_{y\in\mathcal{Y}_c} \sigma_\alpha(\hat{r}(x)^\top\hat{r}(y)),$$

$$\sigma_\alpha(u) = \begin{cases}\mathrm{sign}(u)|u|^\alpha & \text{if } u>\tau\\ 0 & \text{otherwise,}\end{cases} \qquad \hat{r}(x) = \frac{x-q(x)}{\|x-q(x)\|}.$$

Aggregated kernel

$$\mathrm{M_A}(\mathcal{X}_c,\mathcal{Y}_c) = \sigma\left\{\psi\left(\sum_{x\in\mathcal{X}_c}\phi(x)\right)^\top \psi\left(\sum_{y\in\mathcal{Y}_c}\phi(y)\right)\right\} \quad (15)$$

$$= \sigma\left(\Phi(\mathcal{X}_c)^\top\Phi(\mathcal{Y}_c)\right), \quad (16)$$

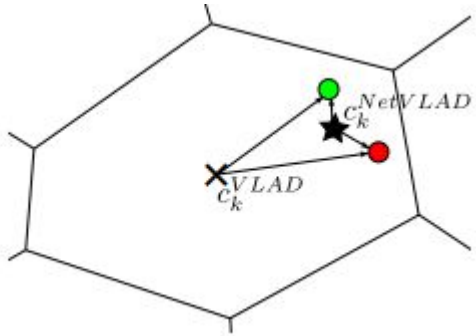
$$\mathrm{ASMK}(\mathcal{X}_c,\mathcal{Y}_c) = \sigma_\alpha\left(\hat{V}(\mathcal{X}_c)^\top\hat{V}(\mathcal{Y}_c)\right)$$

$$\Phi(\mathcal{X}_c) = \hat{V}(\mathcal{X}_c) = V(\mathcal{X}_c)/\|V(\mathcal{X}_c)\|$$

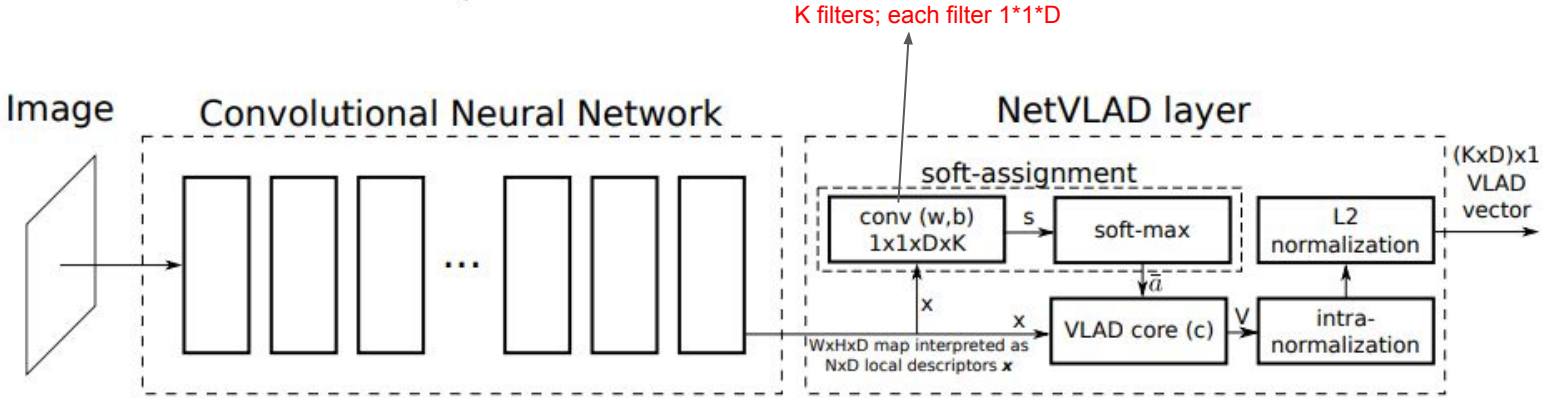
Tolias, et al.., Image Search with selective match kernals: aggregation across single and multiple images, IJCV 2015

Code

NetVLAD

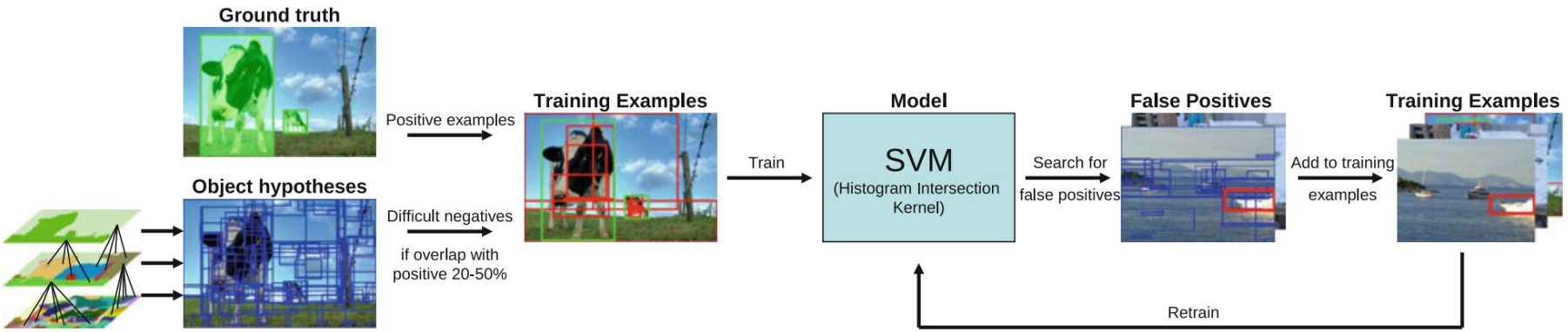


- Contributions
  - What is a good CNN architecture for place recognition? mid-level (Conv5) convolutional features; NetVLAD; PCA
  - How to gather sufficient amount of annotated data for the training?
  - How can we train the developed architecture in an end-to-end manner tailored for the place recognition task?



Arandjelovic, et al.., NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016

Selective Search



[OpenCV]

[J.R.R.Uijlings, et al.., Selective Search for Object Recognition, IJCV 2013]

Selective Search

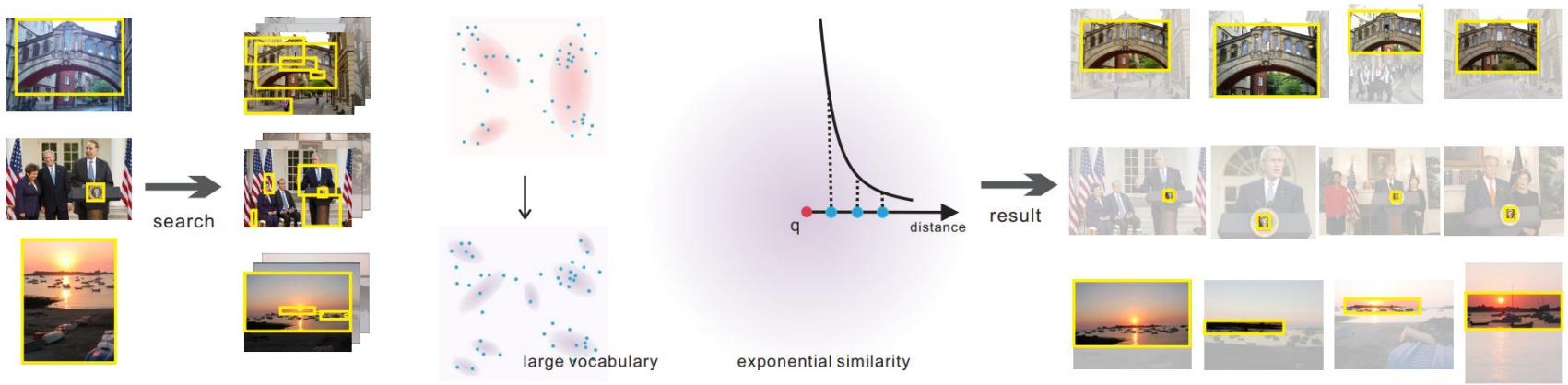


Figure 1. We propose locality in generic instance search from one example. As the first novelty, we consider many boxes as candidate targets to search locally in the picture by an efficient point-indexed representation. The same representation allows, as the second novelty, the application of very large vocabularies in Fisher vector and VLAD to search locally in the feature space. As the third novelty, we propose the exponential similarity to emphasize local matches in feature space. The method does not only improve the accuracy but also delivers a reliable localization.

Ran Tao, et al.., Locality in Generic Instance Search from One Example, CVPR 2014

GeM (generalized-mean pooling)

$$\mathbf{f}^{(g)} = [\mathrm{f}_1^{(g)} \dots \mathrm{f}_k^{(g)} \dots \mathrm{f}_K^{(g)}]^\top, \quad \mathrm{f}_k^{(g)} = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^{p_k} \right)^{\frac{1}{p_k}}.$$

$$\frac{\partial \mathrm{f}_k}{\partial x_i} = \frac{1}{|\mathcal{X}_k|} \mathrm{f}_k^{1-p_k} {x_i}^{p_k - 1},$$

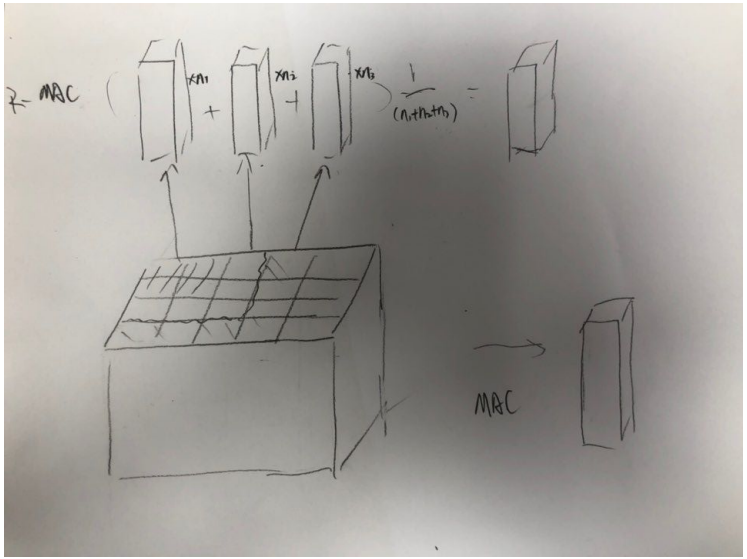
$$\frac{\partial \mathrm{f}_k}{\partial p_k} = \frac{\mathrm{f}_k}{p_k^2} \left(\log \frac{|\mathcal{X}_k|}{\sum_{x \in \mathcal{X}_k} x^{p_k}} + p_k \frac{\sum_{x \in \mathcal{X}_k} x^{p_k} \log x}{\sum_{x \in \mathcal{X}_k} x^{p_k}} \right).$$

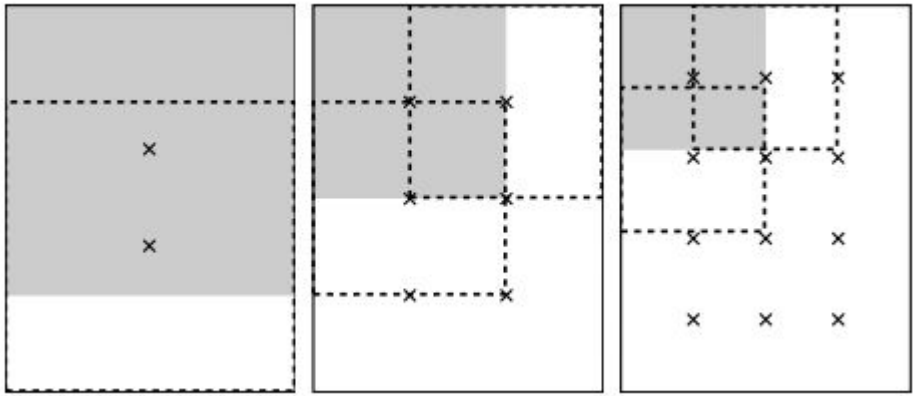
- Contributions
  - Exploit **SfM information** and enforce, not only hard non-matching (negative), but also hard-matching (positive) examples from CNN training.
  - Show that the whitening traditionally performed on short representations is, in some cases, unstable. So propose to learn the whitening through the same training data.
  - Propose a **trainable pooling layer** that generalizes existing popular pooling schemes for CNNs.
  - Propose a novel weighted query expansion that is more robust.
  - A new state-of-the-art result.

Radenovic, et al.., defence

Radenovic, et al.., Fine-tuning CNN Image Retrieval with No Human Annotation, TPAMI 2018

R-MAC





[Tolias, et al... Particular Object Retrieval with Integral Max-pooling of CNN activations. ICLR 2016 (591)]

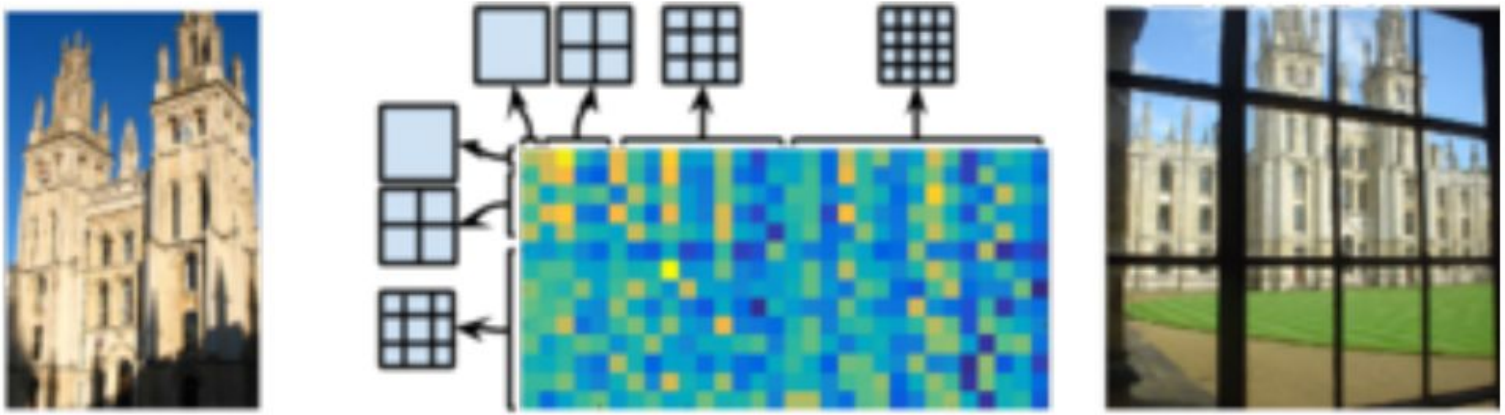
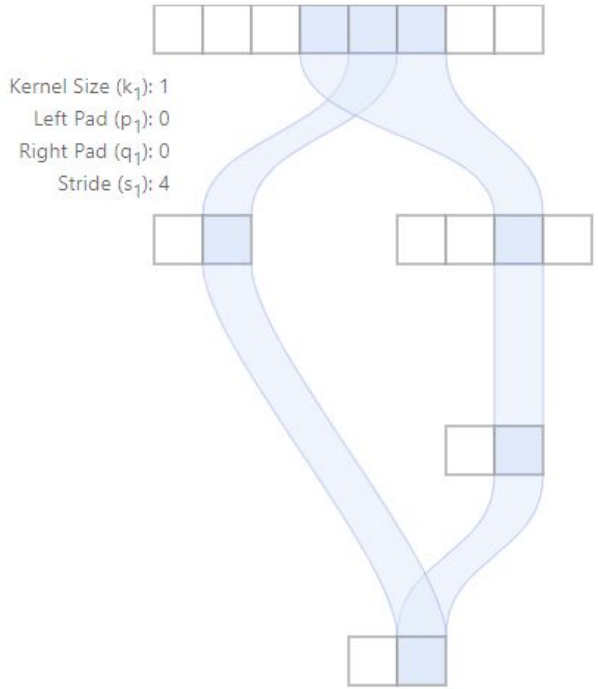


Fig. 5 **Schematic of distance matrix between two images:** The distance between two images is computed by pooling the distances between different patches of the query and the reference image according to equations (3-4).

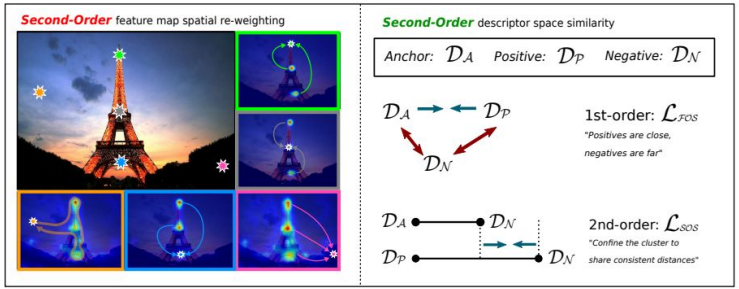
A.S.Razavian, et al.., Visual Instance Retrieval with Deep Convolutional Networks, ITE Transaction on Media Technology and Applications 2016

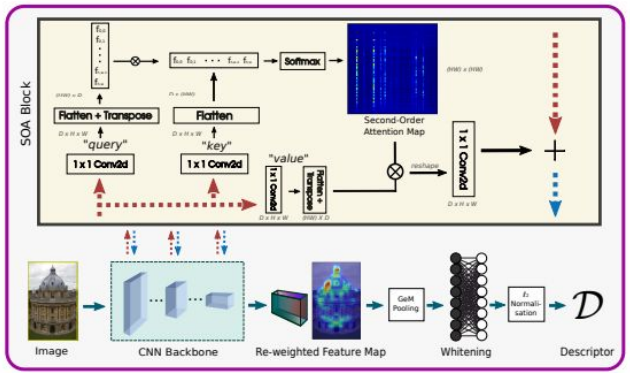
Computing receptive fields

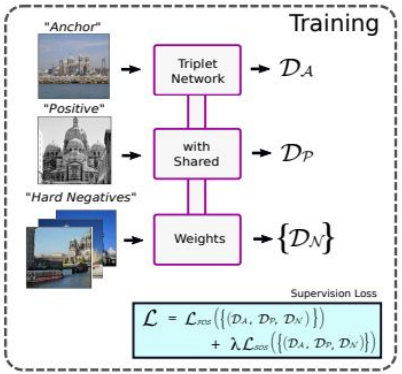


Araujo, et al.., Computing Receptive Fields of Convolutional Neural Networks, Distill 2019

Second-Order Loss

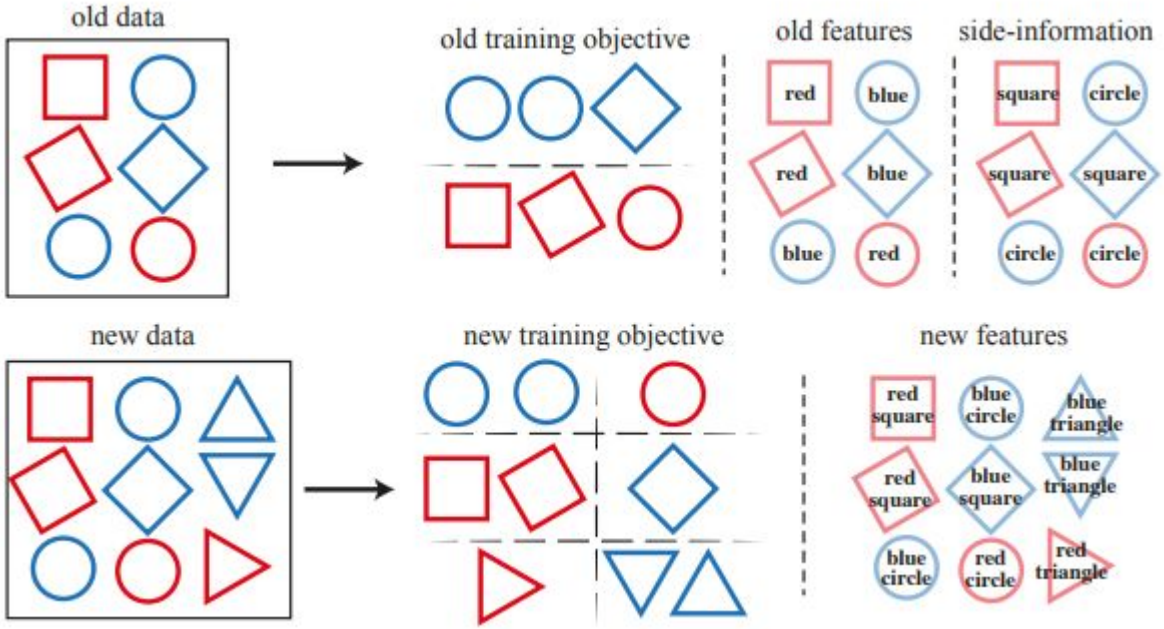






Ng, et al.., SOLAR: Second-Order Loss and Attention for Image Retrieval, ECCV 2020

Forward Compatible Training



Ramanujan, et al.., Forward Compatible Training for Large-Scale Embedding Retrieval Systems, CVPR 2022

Whiten

burstiness

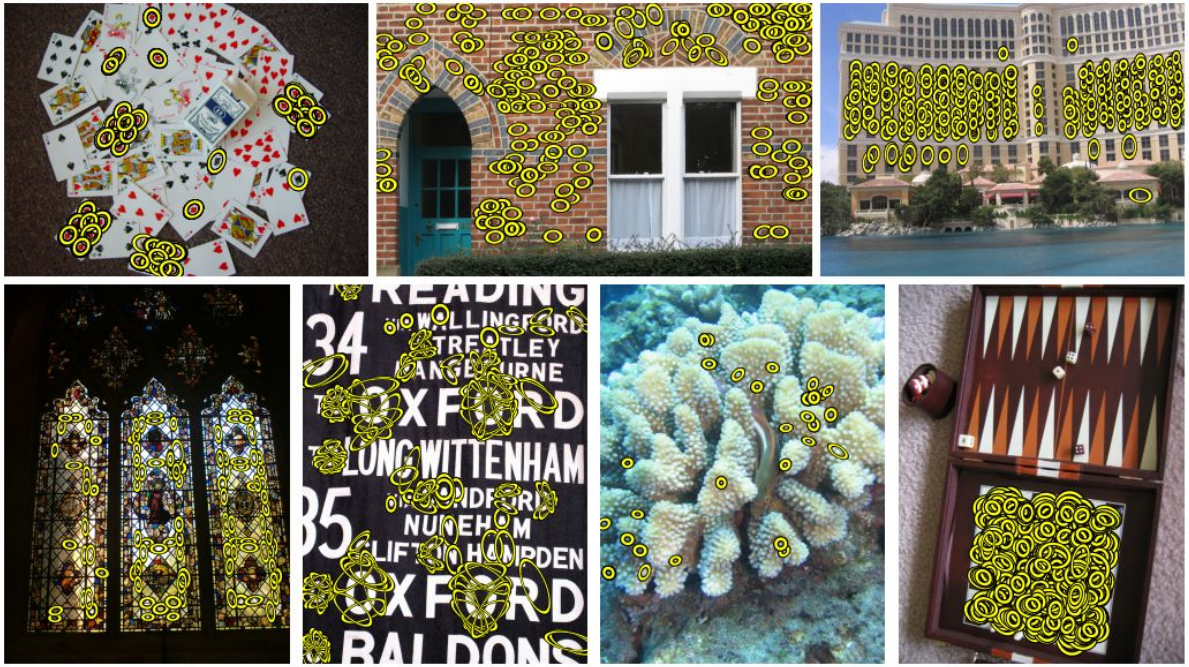


Figure 1. Illustration of burstiness. Features assigned to the most "bursty" visual word of each image are displayed.

Jegou, et al.., On the burstiness of visual elements, CVPR 2009

Co-occurrence

水元素被重复计算

Figure 1. A query image (left) and top ranked images retrieved by two methods from a database of 5 million images: standard (*tf-idf* with spatial verification) retrieval (top) and the same method after automatic detection and removal of co-ocset features (bottom).

Chum, et al.., Unsupervised Discovery of Co-occurrence in Sparse High Dimensional Data, CVPR 2010

Benefit of PCA and whitening: Co-missing and co-occurrence

$$\hat{X} = \frac{\mathrm{diag}(\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_{D'}^{-\frac{1}{2}})\boldsymbol{P}^\top X}{\left\|\mathrm{diag}(\lambda_1^{-\frac{1}{2}}, \ldots, \lambda_{D'}^{-\frac{1}{2}})\boldsymbol{P}^\top X\right\|},$$

Jegou, et al.., Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening, ECCV 2012

Classical Features

Hessian-affine detector

Perdoch, et al.., Efficient repersentation of local geometry for image scale object retrieval, CVPR 2009

rootSIFT descriptor

Arandjelovic, et al.., Three things everyone should know to improve object retrieval, CVPR 2012

CNN Features

5

Faster R-CNN features for Instance Search

- Global Features + Local Features

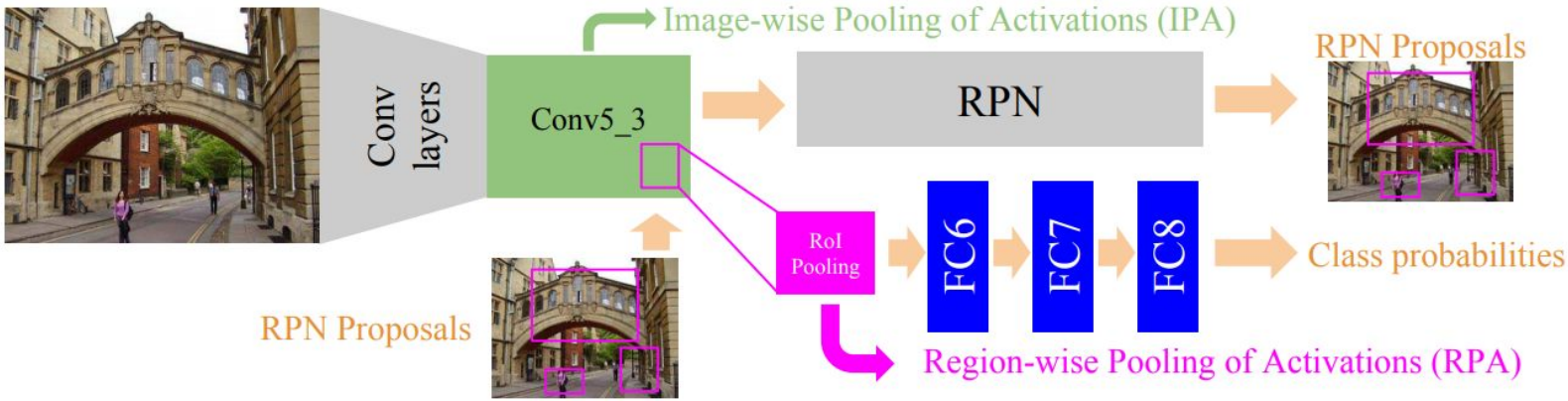


Figure 2. Image- and region-wise descriptor pooling from the Faster R-CNN architecture.

Salvador, et al.., Faster R-CNN Features for Instance Search, CVPR 2016 (Workshop)

Bags of Convolutional Features

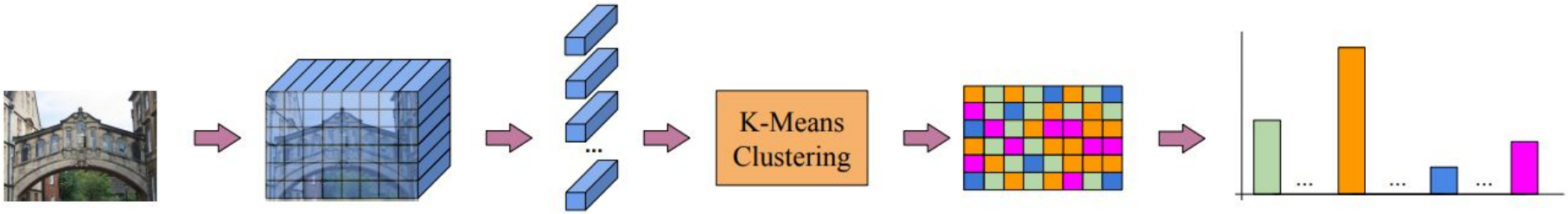
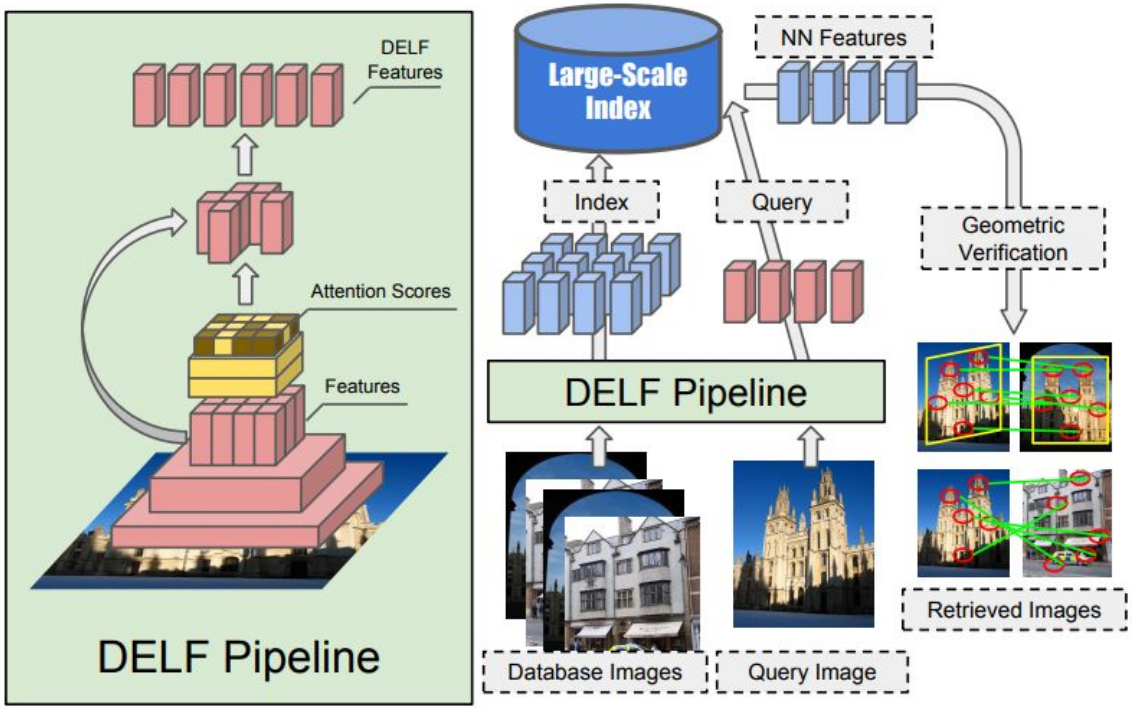


Figure 2: The Bag of Local Convolutional Features pipeline (BLCF).

Salvador, et al.., Bags of Local Convolutional Features for Scalable Instance Search, ICMR 2016 (123)

DELF (deep local features)

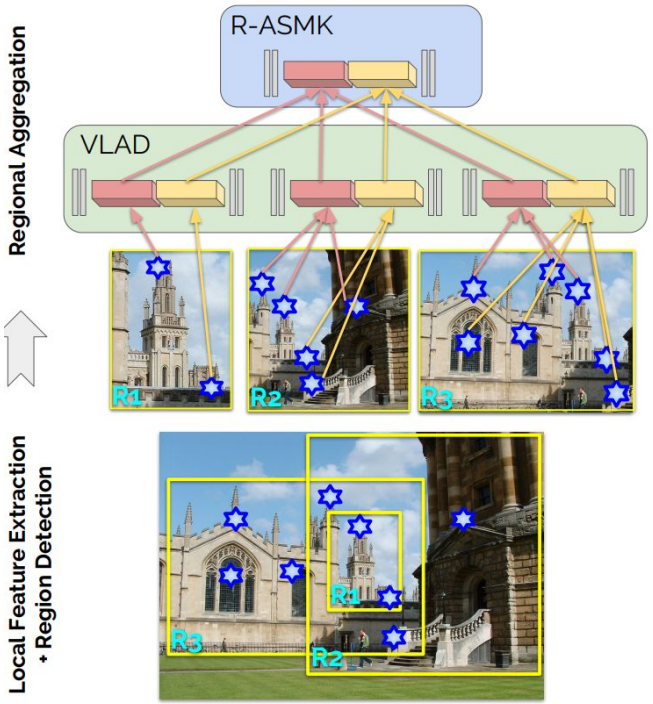


- Contributions
  - A new large-scale dataset, Google-Landmarks, which contains more than 1 M landmark images from almost 13K unique landmarks.
  - Propose DELF, a CNN-based local feature with attention, which is trained with weak supervision using image-level class labels only.

Noh, et al.., Large-scale image retrieval with attentive deep local features, ICCV 2017

Code

Detect-to-Retrieve



$$\text{sim}^{(\text{R-VLAD})}(X, Y^{(n)})$$
$$= \gamma(\mathcal{X}) \sum_c \sum_r \frac{\gamma(\mathcal{Y}^{(n,r)})}{R_n} V(\mathcal{X}_c)^T V(\mathcal{Y}_c^{(n,r)})$$
$$= \sum_c \gamma(\mathcal{X}) V(\mathcal{X}_c)^T \sum_r \frac{\gamma(\mathcal{Y}^{(n,r)})}{R_n} V(\mathcal{Y}_c^{(n,r)})$$
$$= \sum_c V_R(\mathcal{X}_c)^T V_R(\{\mathcal{Y}_c^{(n,r)}\}_r)$$

$$\text{sim}^{(\text{R-ASMK})}(X, Y^{(n)}) =$$
$$\sum_c \sigma\left(\hat{V}_R(\mathcal{X}_c)^T \hat{V}_R(\{\mathcal{Y}_c^{(n,r)}\}_r)\right)$$

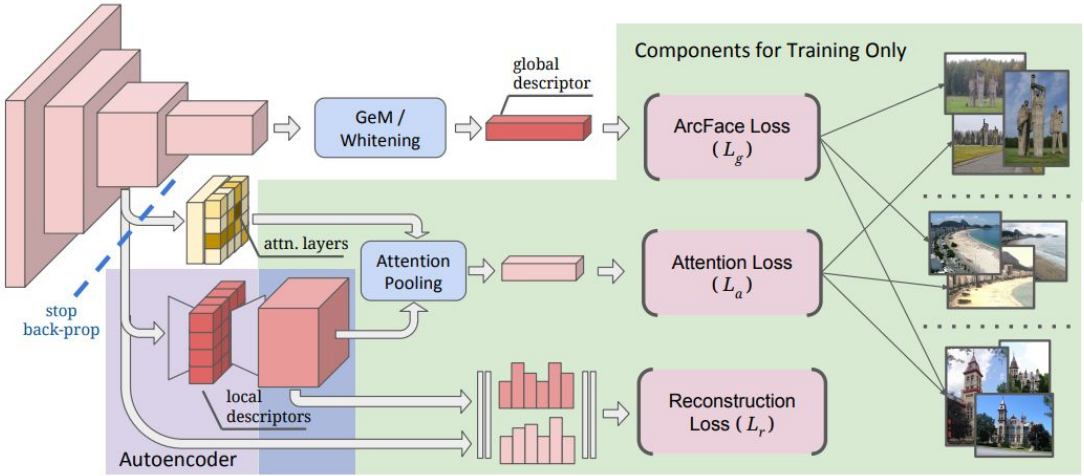
$$\hat{V}_R(\{\mathcal{Y}_c^{(n,r)}\}_r) = \frac{V_R(\{\mathcal{Y}_c^{(n,r)}\}_r)}{\left\|V_R(\{\mathcal{Y}_c^{(n,r)}\}_r)\right\|}$$

- Contributions
  - Propose Google Landmark Boxes datasets (add landmark boxes on Google Landmark dataset).
  - Leverage the trained detector and produce more efficient regional search systems.
  - Propose regional aggregated match kernels to leverage selected image regions and produce a discriminative image representation.

出现次数较少的visual patterns应该更重要

Teichmann, et al.., Detect-to-Retrieve: Efficient Regional Aggregation for Image Search, CVPR 2019

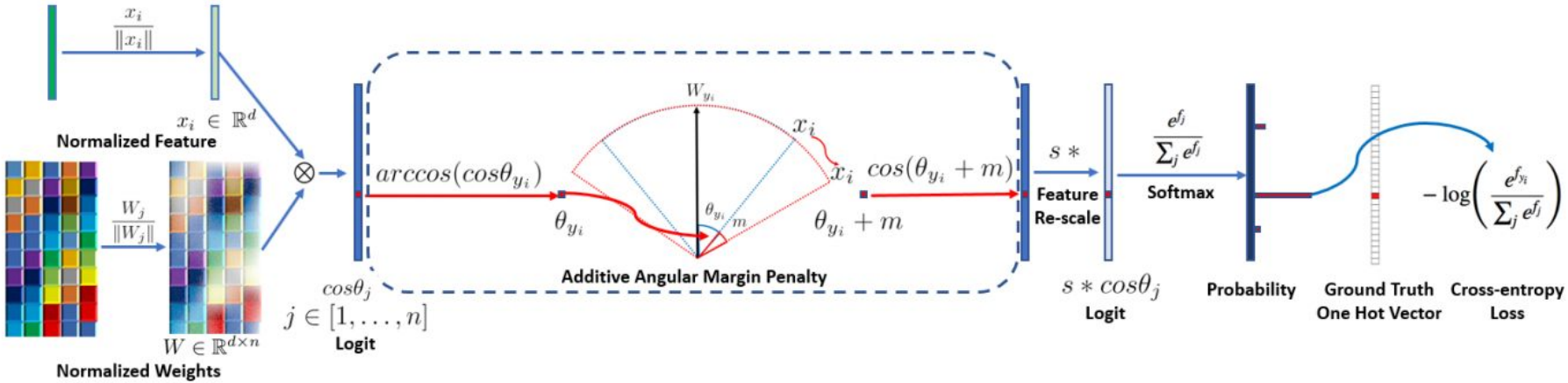
Code
DELG



- Contributions
  - **DELG** (DEep Local and Global features), a unified model to represent both local and global features, using a convolutional neural network (CNN).
  - Adopt a convolutional **autoencoder module** that can successfully learn low-dimensional local descriptors. (avoid the need of post-processing learning steps, such as PCA.)
  - Design a procedure that enables end-to-end training of the proposed model using only image-level supervision. (**control the gradient flow**)

Cao, et al.., Unifying Deep Local and Global Features for Image Search, ECCV 2020

ArcFace



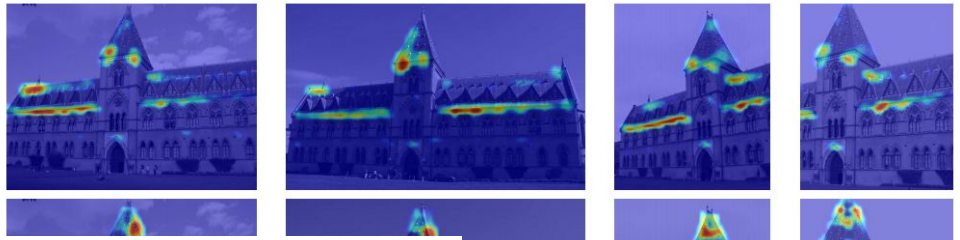
Deng, et al.., ArcFace: Additive Angular Margin Loss for Deep Face Recognition, CVPR 2019

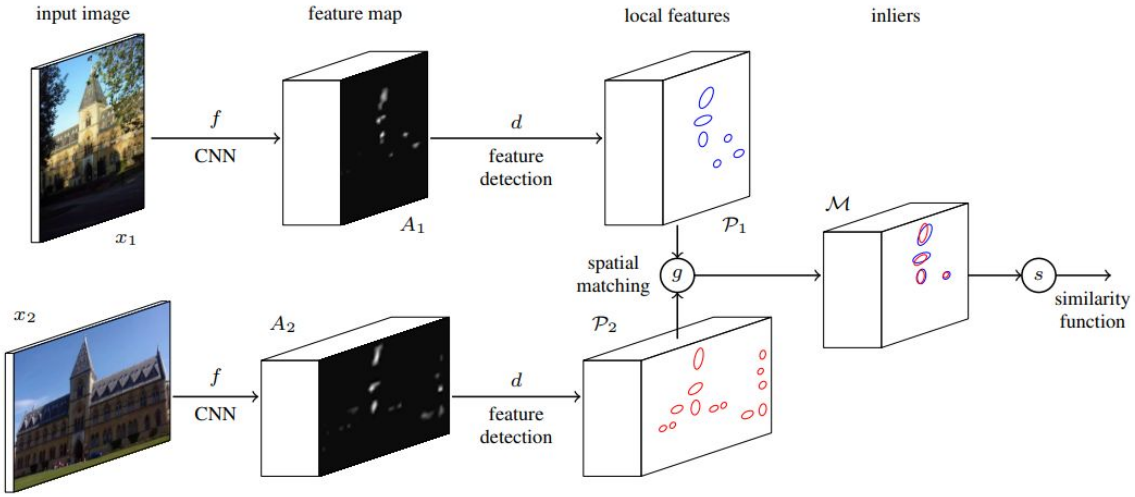
Unsupervised

Fig. 2.
Examples of training query images (green border) and matching images selected as positive examples by methods (from left to right) $m_1(q)$, $m_2(q)$, and $m_3(q)$. (Color figure online)

Radenovic, et al.., CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples, ECCV 2016

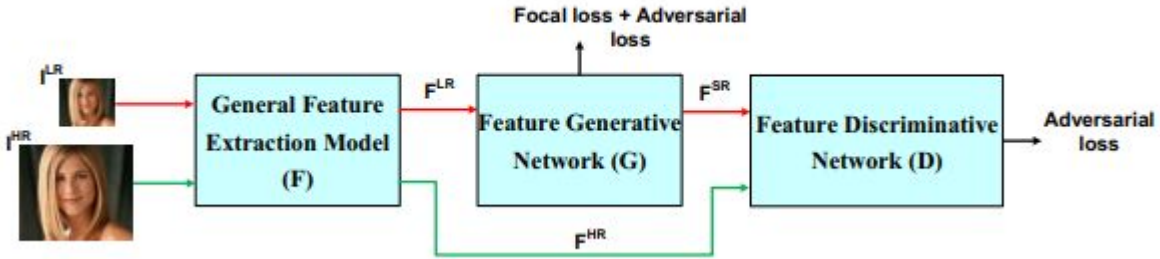
Spatial matching per channel





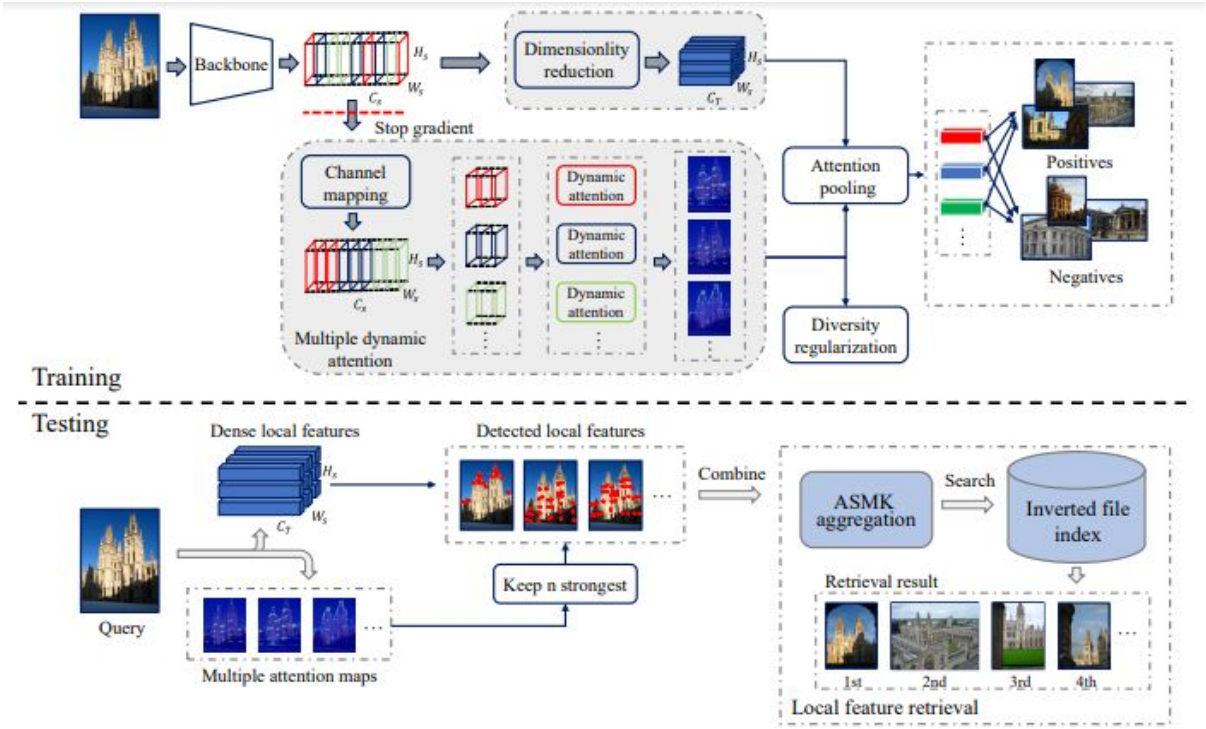
Simeoni, et al.., Local Features and Visual Words Emerge in Activations, CVPR 2019

Feature Super-Resolution



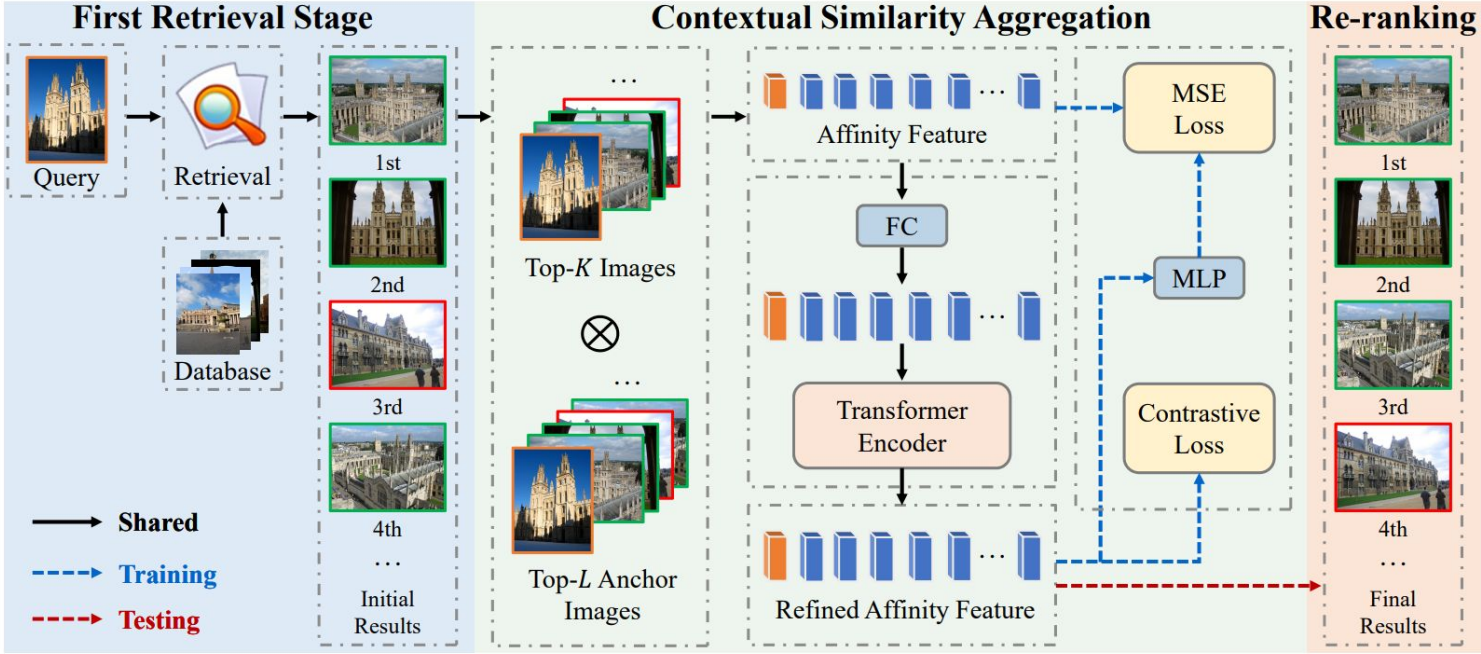
Tan, et al.., Feature Super-Resolution: Make Machine See More Clearly, CVPR 2018

Attention local features



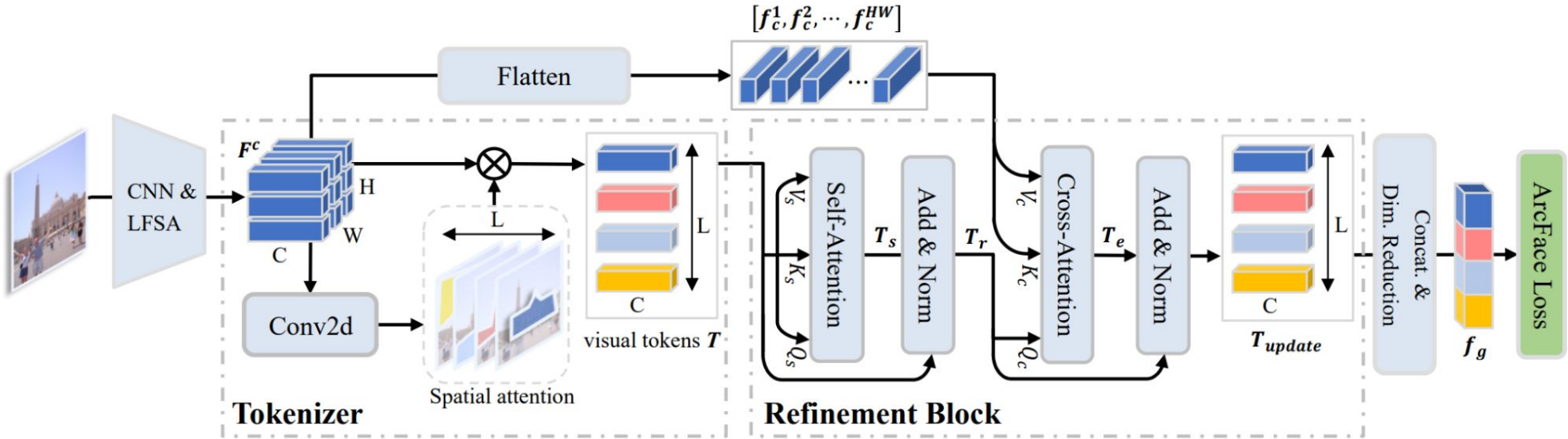
Hui Wu, et al.., Learning Deep Local Features With Multiple Dynamic Attensions for !, ICCV 2021

Transformer + query expansion



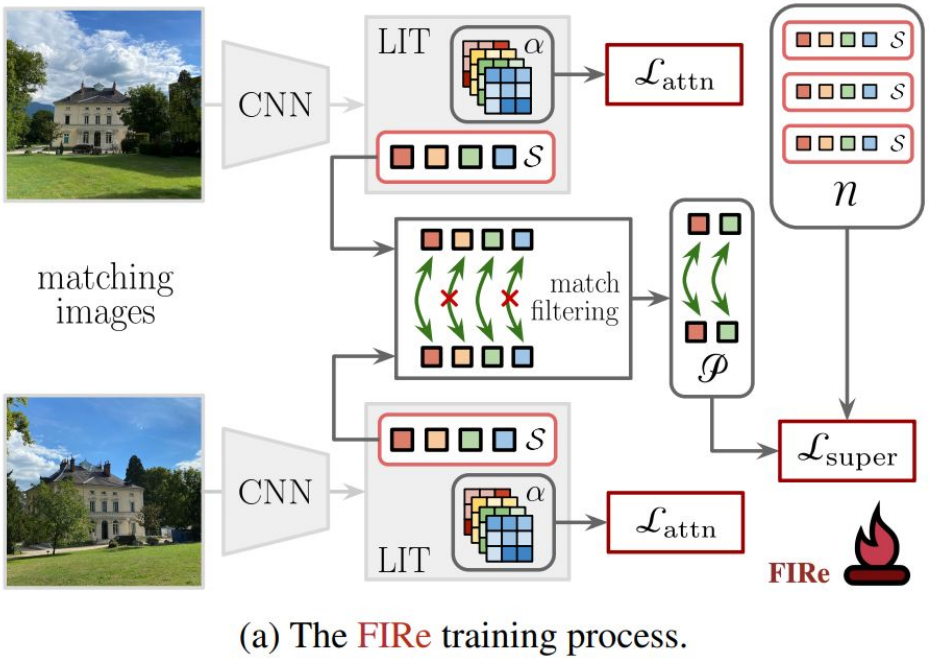
Hui Wu., et al.., Contextual Similarity Aggregation with Self-attention for Visual Re-ranking, NeurIPS2021

Transformer + query expansion



Hui Wu., et al.., Learning Token-based representation for Image Retrieval. AAAI 2022

Contrastive learning + transformer



(a) The FIRe training process.

(b) The LIT module architecture.

Weinzaepfel., et al.., Learning Super Features For Image Retrieval, ICLR2022

Sub-image search

2

Object-based image retrieval

- The combination of content based image retrieval + object localization
  - Give a object patch, how to find the same object in the database.
  - High accuracy + large set of candidate images.

Figure 1. *Object-based image retrieval.* Left: The users marks an object in an image by a bounding region. Right: The system returns images from a database that show the query object.

Lampert, et al.., Detecting objects in large image collections and videos by efficient subimage retrieval, ICCV 2009 (80)

A local Bag-of-Features Model

Represent each database images as a family of histograms that depend functionally on a bounding rectangle.

Fig. 1. An example of small object retrieval. Left: a query image with a region of interest. Right: the top 8 retrieved images using our approach and the baseline Global BoF approach. Red rectangle represents the query object of interest, and Green rectangles represent the returned object bounding boxes using our approach. In contrast, the baseline method cannot return bounding boxes and need additional totally different criterion (*e.g.* RANSAC) to localize objects.

Lin, et al.., A Local Bag-of-Features Model for Large-Scale Object Retrieval, ECCV 2010

Unsupervised Object discovery

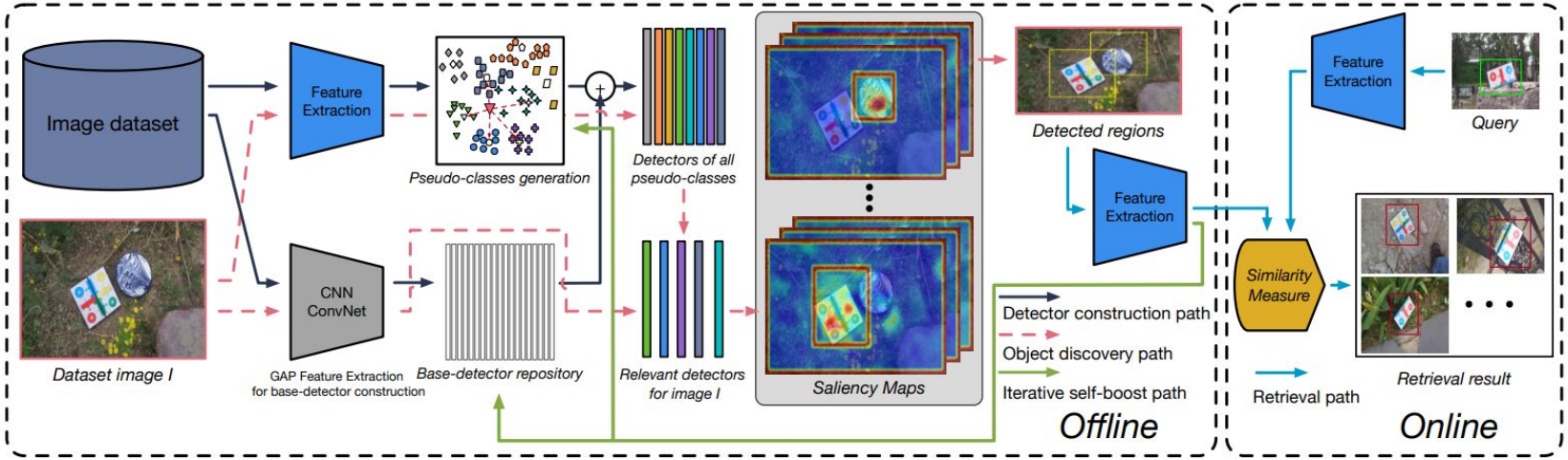


Fig. 1. Overview of the proposed dataset-driven unsupervised object discovery framework (DUODIS in short) for region-based instance image retrieval. Our framework consists of two (i.e., offline and online) parts. In the offline part, there are three kinds of paths. The detector construction path (in black) exploits the whole image dataset and constructs the base-detector repository and the object detectors for all pseudo-classes. The object discovery path (in red and dashed line) assigns relevant object detectors to each image and applies them to conduct object discovery. The iterative self-boosting path (in green) utilizes the detected regions to further enhance the whole framework. In the online part, once a query object is submitted, its features will be extracted to measure its similarity to each of the regions detected from the image dataset. The most similar images are then retrieved, with the matched object regions delineated. In instance image retrieval, it is usually assumed that a user will indicate the object to search for in a query image by a bounding box. This setting is followed in this paper.

Zhang, et al..,Dataset-driven Unsupervised Object Discovery for Region-based Instance Image Retrieval TPAMI 2022

Grid-based

Fig. 1. *An example of visual object search.* Left: a query object, such as a logo, selected by the user. Right: resulting images from visual object search, where object locations are identified and marked by blue bounding boxes.

Jiang, et al... GRID-BASED LOCAL FEATURE BUNDLING FOR EFFICIENT OBJECT SEARCH AND LOCALIZATION , ICIP 2011

Branch and Bound

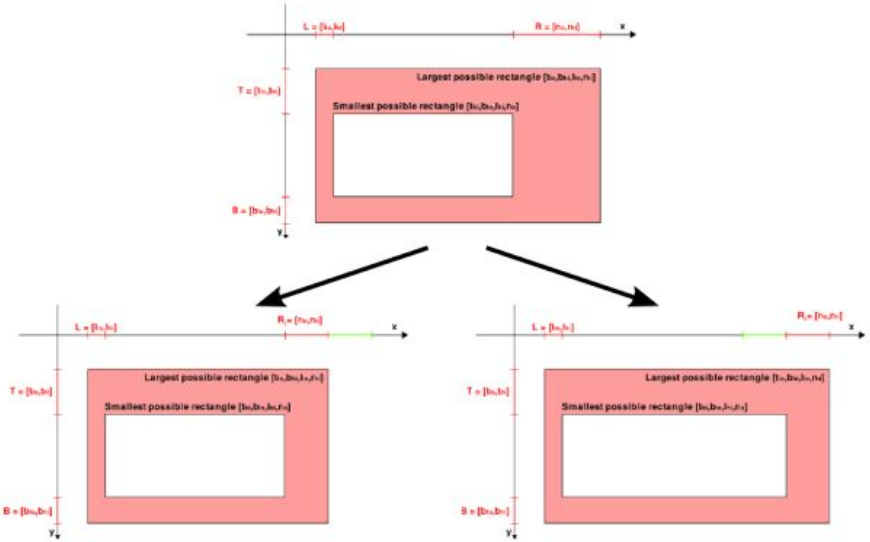


Fig. 2. Splitting rectangle sets is done by dividing one of the intervals in two. In this case, $[T, B, L, R] \rightarrow [T, B, L, R_1] \dot{\cup} [T, B, L, R_2]$, where $R_1 := [r_{lo}, \lfloor \frac{r_{lo}+r_{hi}}{2} \rfloor]$ and $R_2 := [\lfloor \frac{r_{lo}+r_{hi}}{2} \rfloor + 1, r_{hi}]$.

Fig. 3. Example images of `cat` (top) and `dog` (bottom) categories of PASCAL VOC 2006 dataset. Objects occur in different sizes and poses, and multiple object instances are possible within one image. Objects are also frequently occluded or truncated.

Lampert, et al.., Efficient Subwindow Search: A Branch and Bound Framework for Object Localization,TPAMI 2009

Strain estimation

Fig 2: Specimen fixed in UTM

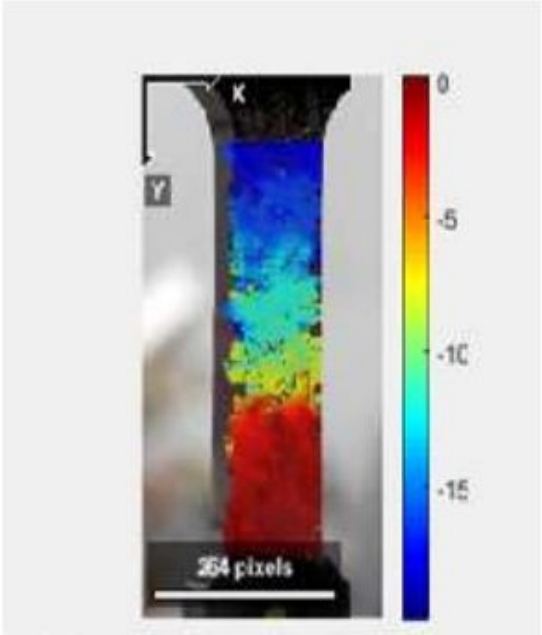


Fig. 5. Strain maps curves obtained on aluminium specimen

Shifani et al...Non contact Two Dimensional Strain Measurement Based on Improved Sub Image Search Algorithm using Correlation Technique , ICSSIT 2019

digital pathology require sub-image retrieval rather than the whole image retrieval for the system to be of **clinical use**

Abstract:

Content-based image retrieval systems for digital pathology require sub-image retrieval rather than the whole image retrieval for the system to be of clinical use. Digital pathology images are huge in size and thus the pathologist is interested in retrieving specific structures from the whole images in the database along with the previous diagnosis of the retrieved sub-image. We propose a content-based sub-image retrieval system (sCBIR) framework for high resolution digital pathology images. We utilize scale-invariant feature extraction and present an efficient and robust searching mechanism for indexing the images as well as for query execution of sub-image retrieval. We present a working sCBIR system and show results of testing our system on a set of queries for specific structures of interest for pathologists in clinical use. The outcomes of the sCBIR system are compared to manual search and there is an 80% match in the top five searches.

Eva Breznik et al..,Content based sub-image retrieval system for high resolution pathology images using salient interest points

, 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society

improve diagnosis and discover patterns in pathologies

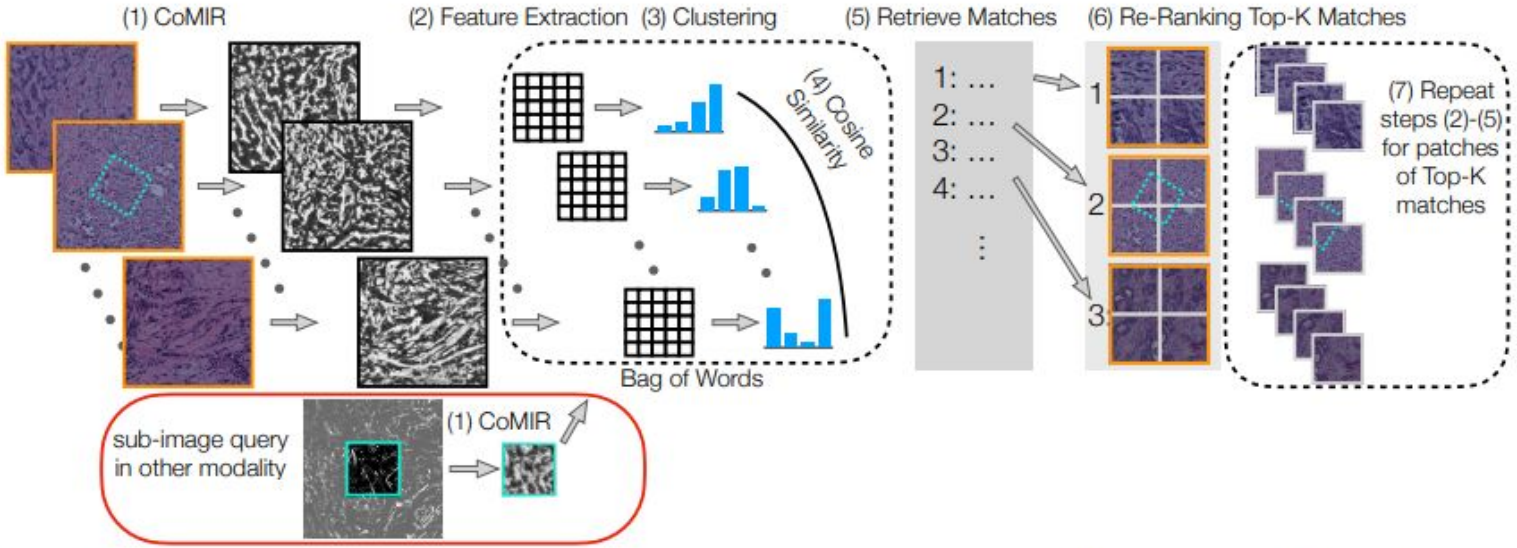


Figure 1: The proposed s-CBIR pipeline learns CoMIRs for the images in the repository and the query in form of a patch, extracts sparse features, which are binned into single descriptors for each image, building the vocabulary for a BoW. Matches are found using the cosine similarity. The Top-K matches are split into patches and a new BoW is computed for Re-Ranking.

Eva Breznik et al.., Cross-Modality Sub-Image Retrieval using Contrastive Multimodal Image Representations , arXiv 2021

Bio

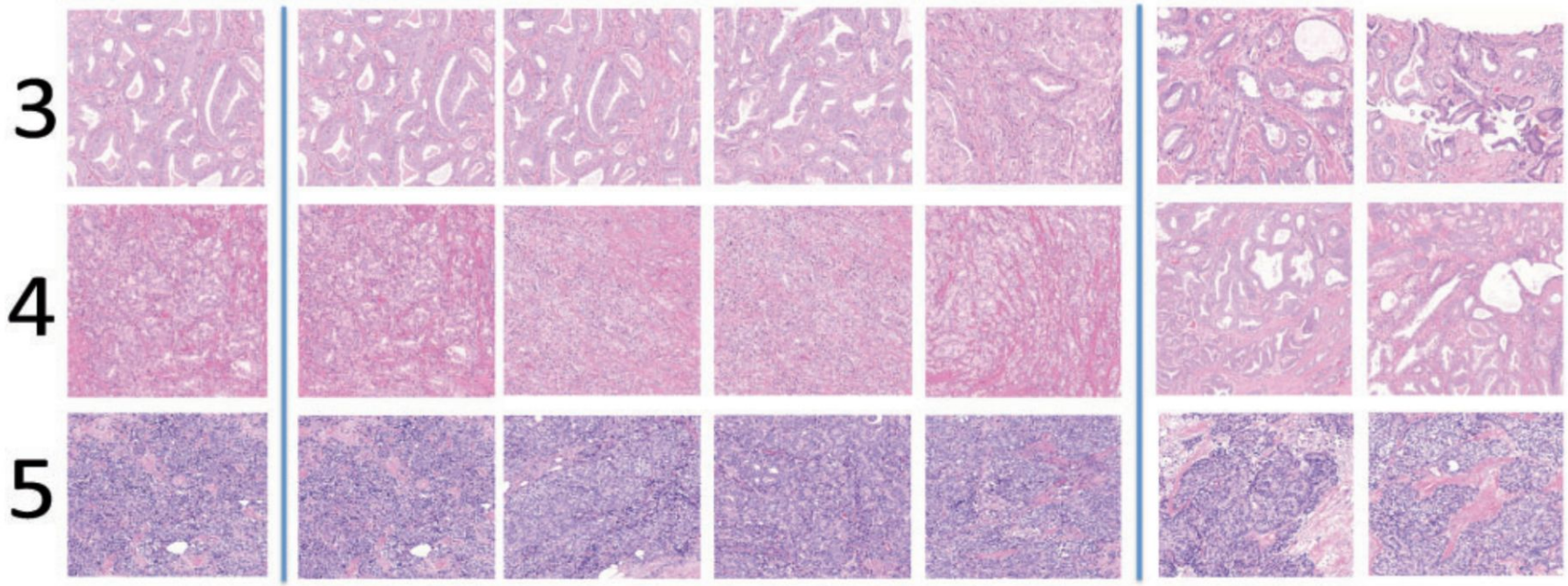
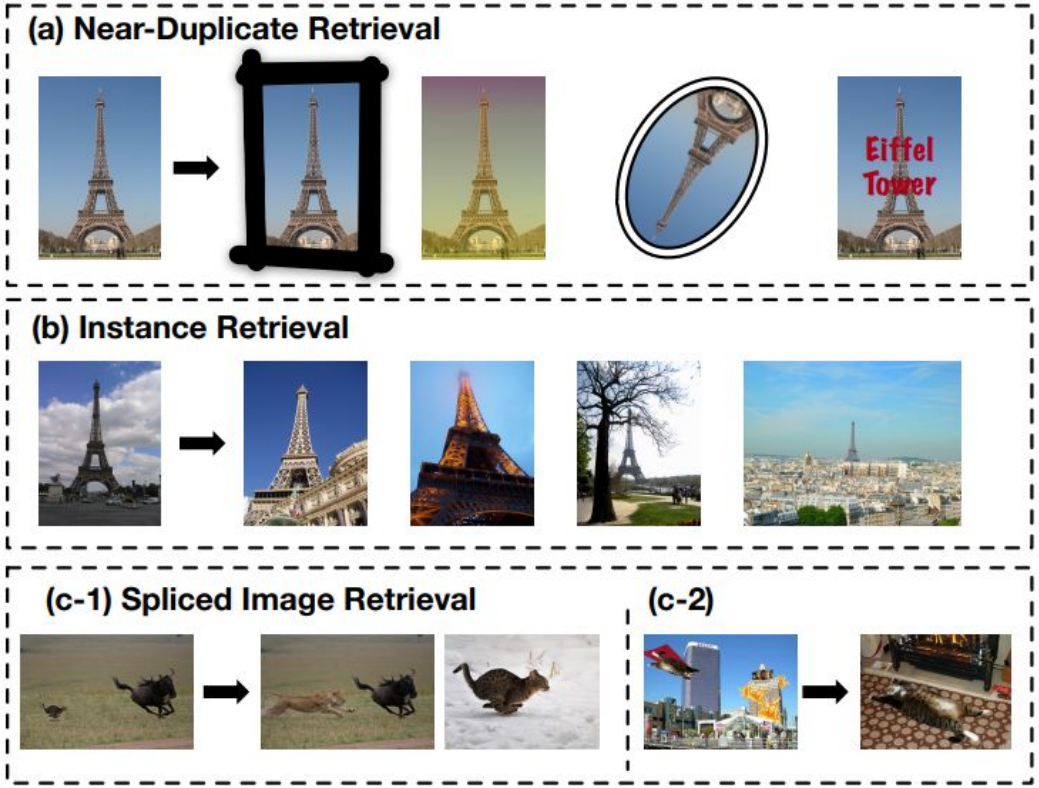


Fig. 4. The sub-image content-based rank retrieval results. The left panel is the query image; the middle panel is the retrieval results representing the most similar cases. The right panel is the retrieval results representing the most dissimilar cases. The number 3, 4 and 5 correspond to different Gleason scores

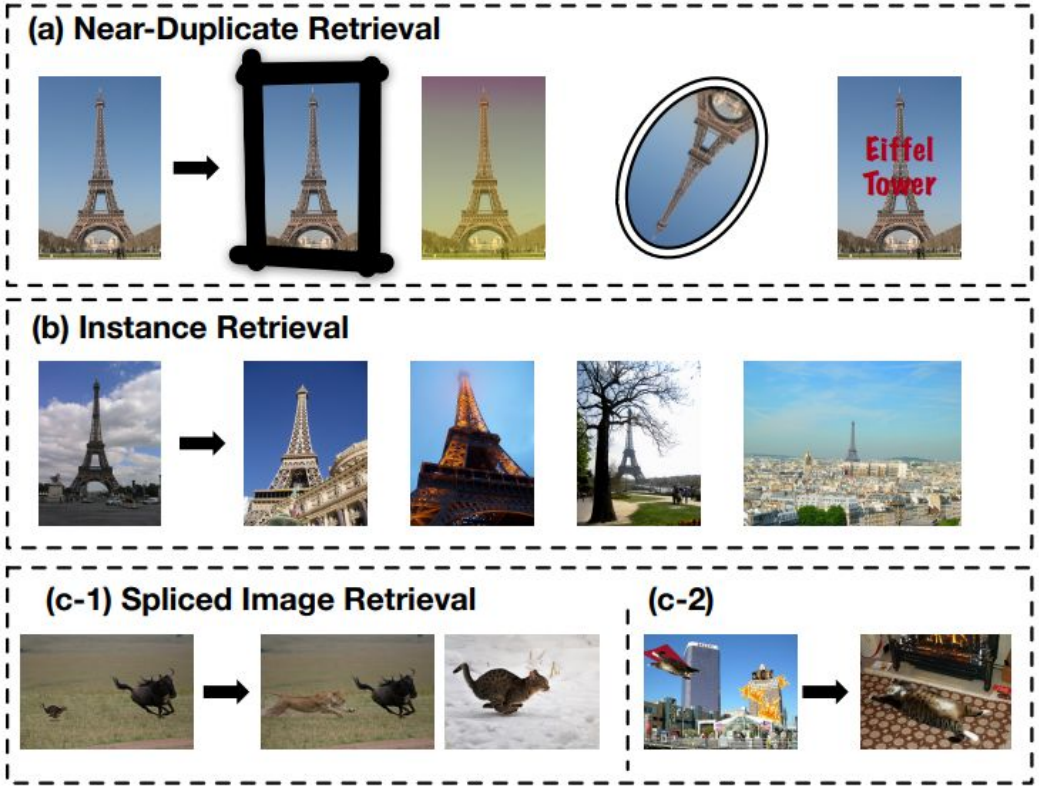
Yang, et al.., Parallel content-based sub-image retrieval using hierarchical searching, Bioinformatics 2014

Spliced Image Retrieval



Chen, et al.., Efficient Object Embedding for Spliced Image Retrieval, CVPR 2021

Spliced Image Retrieval

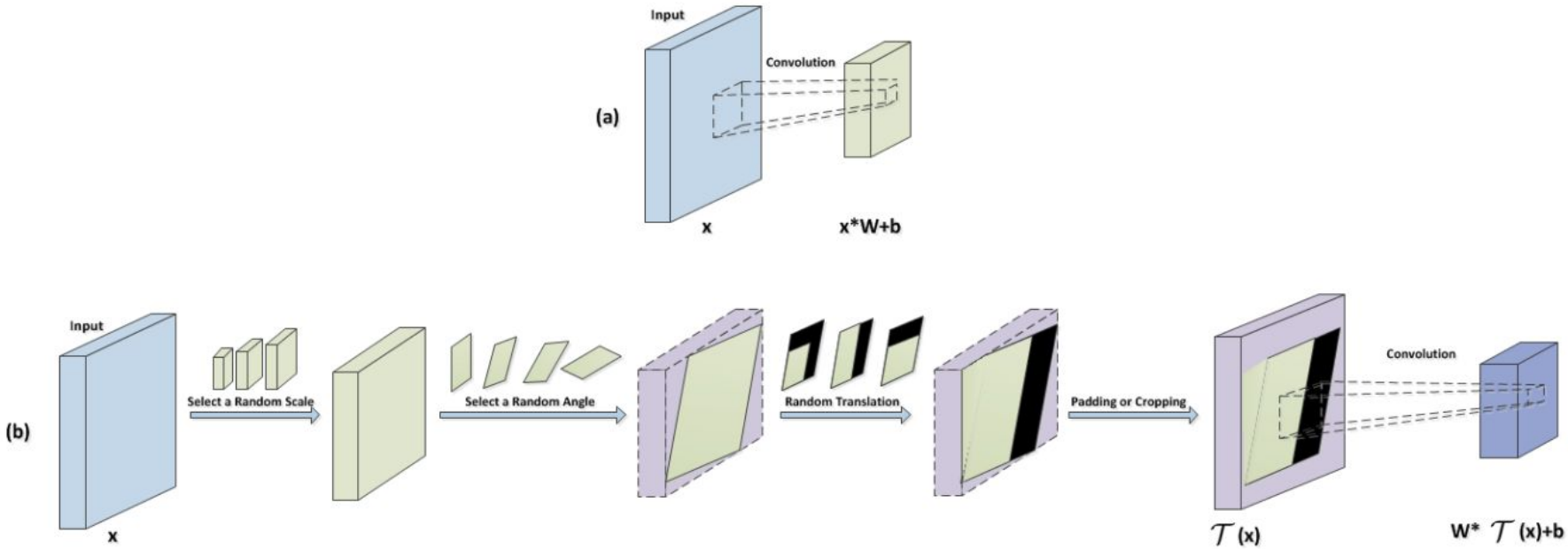


Chen, et al.., Efficient Object Embedding for Spliced Image Retrieval, CVPR 2021

Transform-Invariant

1

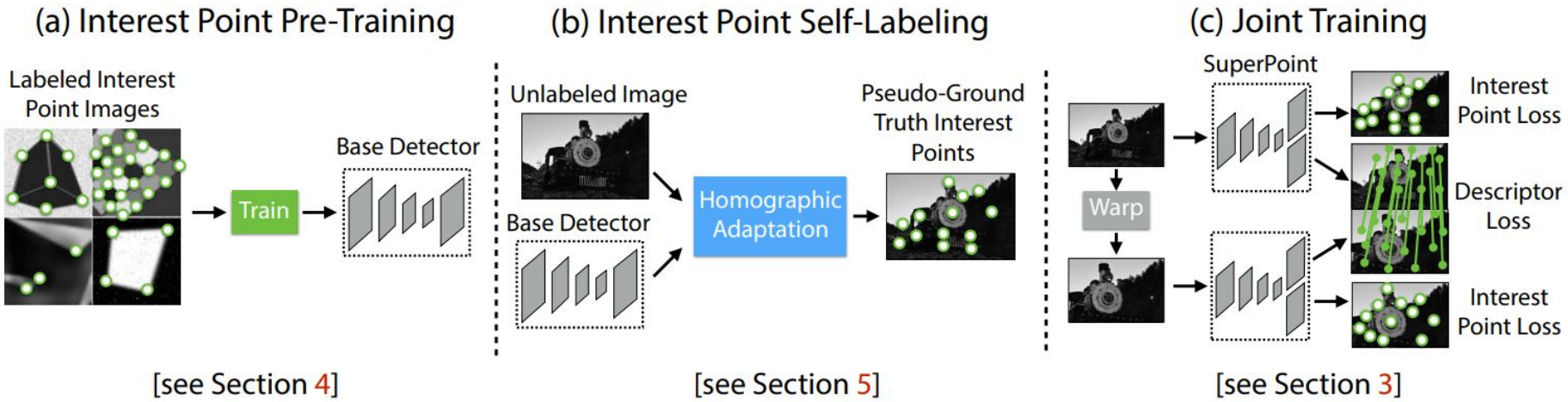
Transformed Invariant Search



Shen, et al..,Transform-Invariant Convolutional Neural Networks for Image Classification and Search, MM 2016

Correspondence features
&
Interest Point

SuperPoint



DeTone, et al.., SuperPoint: Self-supervised Interest Point Detection and Description, CVPR 2018

Dense Pixel Matching

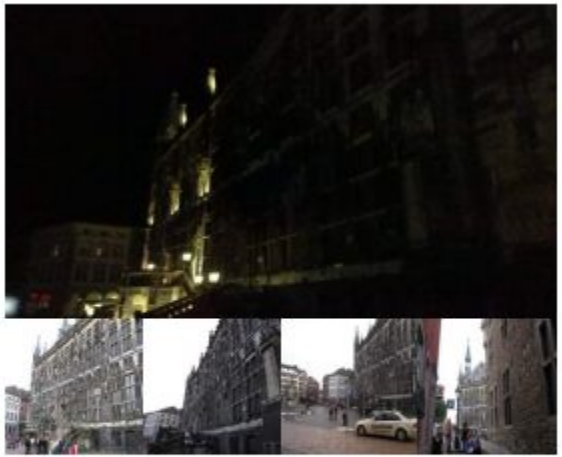


Figure 1: Qualitative results of the proposed method for the task of image retrieval. The first row is a query taken at night-time with a mobile phone camera and the last row is a list of top-4 retrieved database images obtained by our method. All 4 are correct matches.

Laskar, et al.., Geometric Image Correspondence Verification by Dense Pixel Matching, WACV 2020

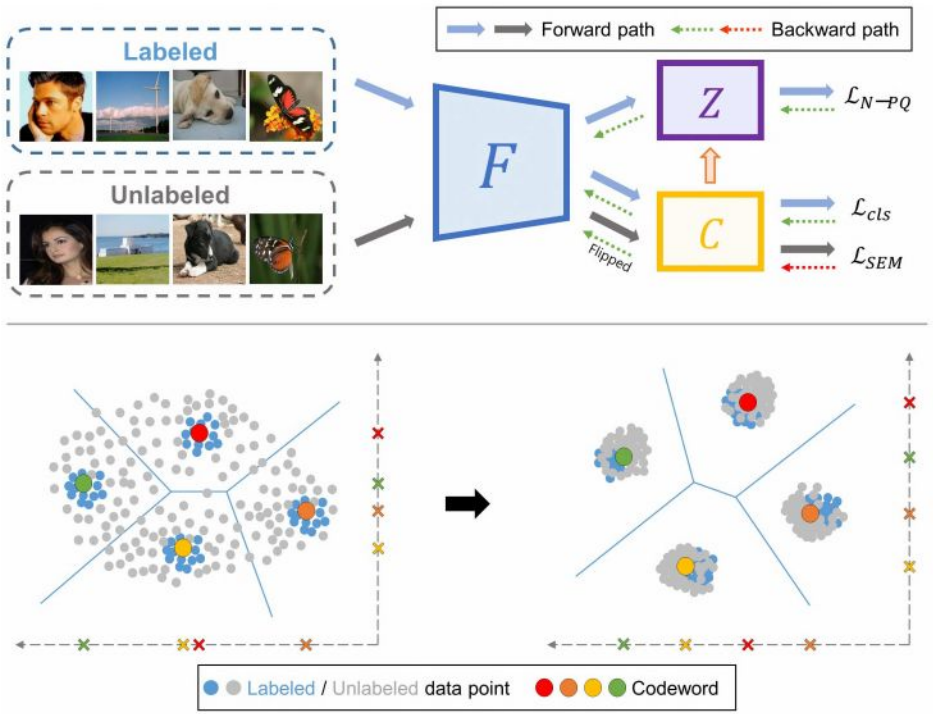
Yi, et al.., Learning to Find Good Correspondences, CVPR 2018

Data Compression

1

Code

GPQ (Generalized Product Quantization)



- Contributions
  - The first deep semi-supervised PQ scheme for image retrieval.
  - With the proposed **metric learning** strategy and **entropy regularization term**, the semantic similarity of labeled data is well preserved into the codewords, and the underlying structure of unlabeled data can be fully used to generalize the network.
  - Yield the state-of-art retrieval results in semi-supervised image retrieval protocols.

Jang, et al.., Generalized Product Quantization Network for Semi-supervised Image Retrieval, CVPR 2020

Feature Visualization

pytorch_cnn_visualization

Visualizing and Understanding Convolutional Networks

Zeilei, et al.., Visualizing and Understanding Convolutional Networks, In: Proc. ECCV 2014 (9898)

Code

Feature visualization

Olah,..., et al.., Feature Visualization, Distill 2017

Code

Computing receptive fields

Araujo,..., et al.., Computing Receiptive Fields of Convolutional Neural Networks, Distill 2019

Correlation Filters

0

With RL; RNN

Mueller, et al.., Context-aware correlation filter tracking, CVPR 2017 (353)

Image clustering

2

simHash for identical image detection

A	B	C	D	E	F	A B C	B C D	A E F
3	6	2	5	4	1	2	2	1
1	2	6	3	5	4	1	2	1
3	2	1	6	4	5	1	1	3
4	3	5	6	1	2	3	3	1

permutations

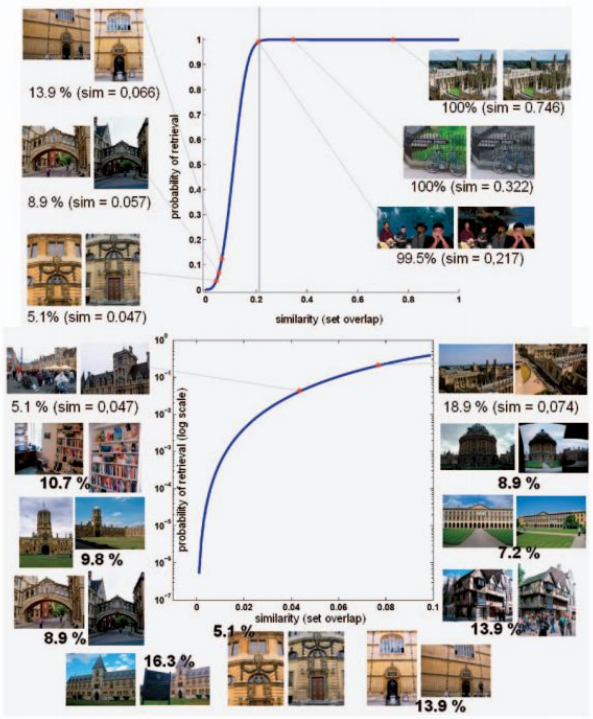
min-hashes

$$P(\min \pi(\mathcal{A}_1) = \min \pi(\mathcal{A}_2)) = \frac{|\mathcal{A}_1 \cap \mathcal{A}_2|}{|\mathcal{A}_1 \cup \mathcal{A}_2|} = \mathrm{sim}(\mathcal{A}_1, \mathcal{A}_2).$$

$$P(\mathcal{A}_1 \approx^m \mathcal{A}_2) = \sum_{i=m}^{k} \binom{k}{i} p^{in} (1 - p^n)^{k-i},$$

Chum, et al.., Scalable Near Identical Image and Shot Detection, CIVR 2007

min-Hash Image Clustering (MHIC)



1) **Hashing.** Image descriptors are stored in a hash table. In our experiments, 2^{51} different descriptor values are used. The probability of two images falling into the same bin (*exact* descriptor match) is proportional to their similarity – eqn. (2).
2) **Similarity estimation.** For all $\binom{n}{2}$ pairs of the n images hashed in the same bin, i.e. for n collisions in the bin, similarity is calculated. The process is fast and consists of comparing two vectors and counting the number of identical elements. In our experiments, the number of vector elements is 512. The similarity is then thresholded.
3) **Spatial consistency.** For each image pair that passed the similarity test, spatial consistency is verified. Image pairs that pass the spatial consistency test become *cluster seeds*.
4) **Seed growing.** Seed images are used as visual queries and the query expansion technique is used to 'crawl' the images in the cluster.

Fig. 2. The min-Hash Image Clustering (MHIC) algorithm.

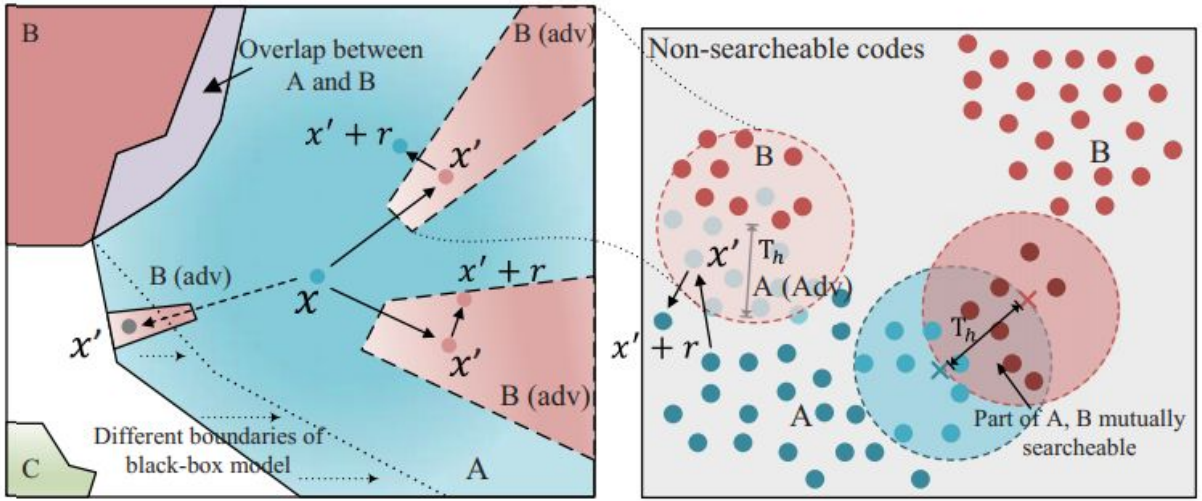
Chum, et al.., Large-scale discovery of spatially related images, TPAMI 2010

Weyand, et al.., Discovering details and scene structure with hierarchical Iconoid shift, ICCV 2013

Philbin, et al.., Geometric latent Dirichlet allocation on a matching graph for large-scale image datasets, IJCV 2011

Attack

You See What I Want You to See



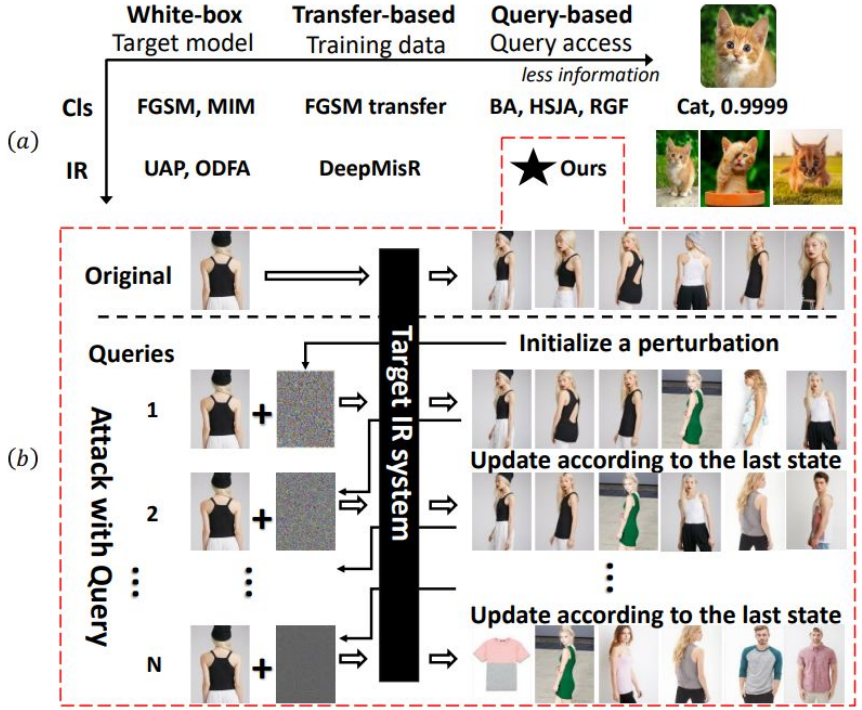
(a) Adv in the pixel space

(b) Adv in the hash space

Figure 4: Illustration of adversarial examples (a) image pixel space (b) hash space.

Xiao, et al..,You See What I Want You to See: Exploring Targeted Black-Box Transferability Attack for Hash-based Image Retrieval Systems, CVPR 2021

Query-efficient attacks

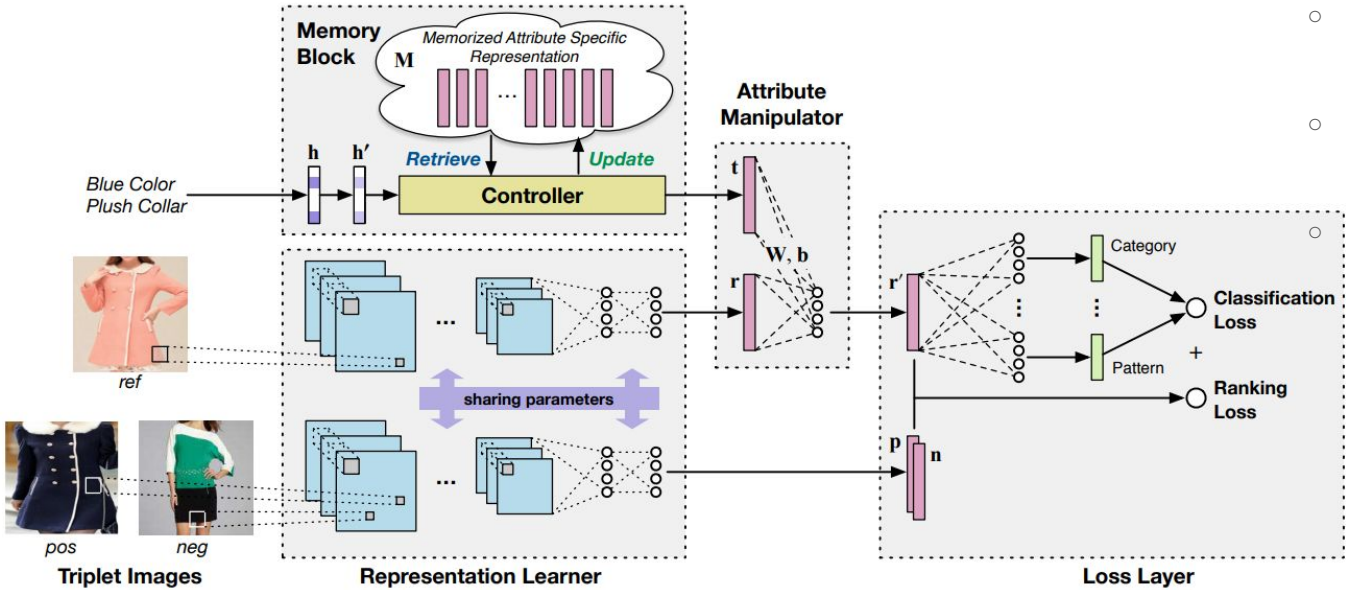


Li, et al..,QAIR: Practical Query-efficient Black-Box Attacks for Image Retrieval, CVPR 2021

Fashion Search

2

Memory-augmented; Interactive Fashion Search

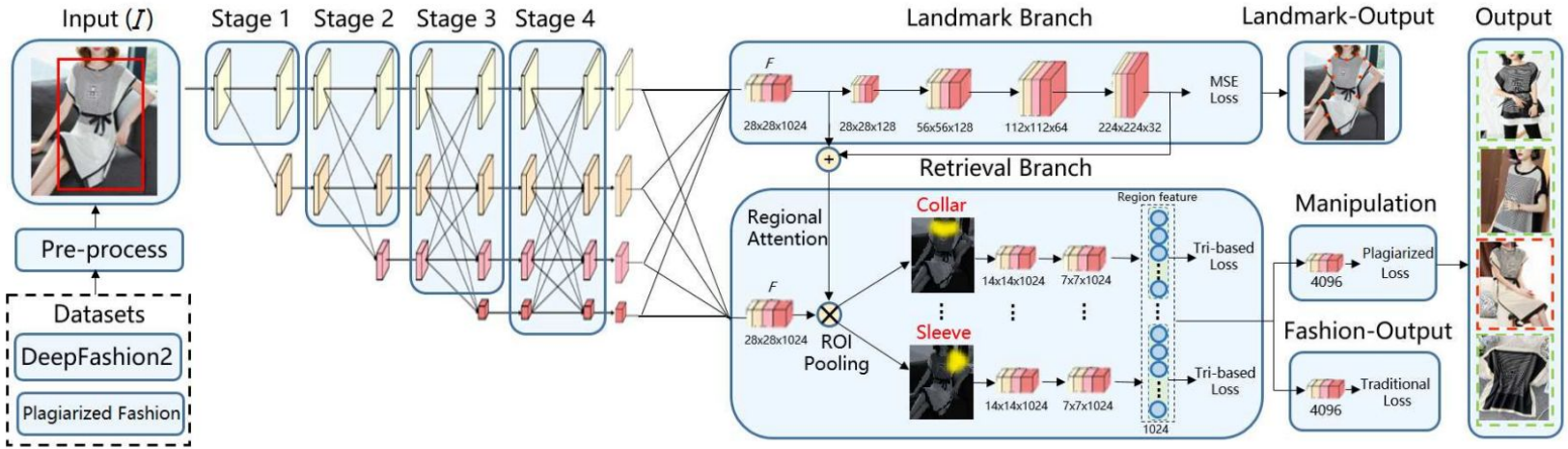


- Contributions
  - Propose a novel memory-augmented attribute manipulation network for interactive fashion search.
  - A memory block consisting of a memory and a neural controller in the proposed networks.
  - Exploit a joint optimization method for attribute representation and retrieval feature learning.

Zhao, et al., Memory-Augmented Attribute Manipulation Networks for Interactive Fashion Search, CVPR 2017

Plagiarized-Search-Net (PS-Net)

- Contributions
  - Proposed PS-Net, which obtain the regional representation of images and compute the similarity region by region
  - A novel problem and a dataset.
  - PS-Net can also be used for traditional fashion retrieval and landmark estimation tasks.



Lang, et al.., Which is plagiarism: Fashion Image Retrieval based on Regional Representation for Design Protection, CVPR 2020

Content-style + text feedback

(a) Example of a **content** modification

(b) Example of a **style** modification

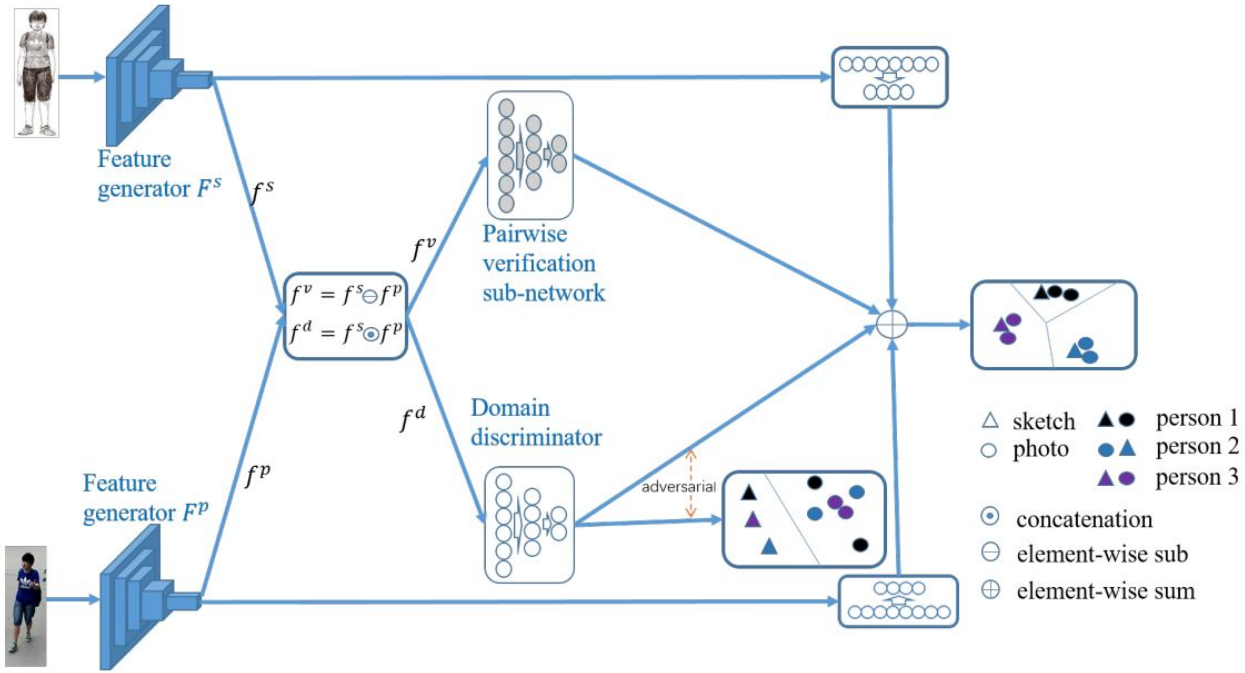
Lee, et al.., CoSMo: Content-Style Modulation for Image Retrieval with Text Feedback, CVPR 2021

Sketch; Re-ID

Cross-domain

4

Sketch-photo person re-identification



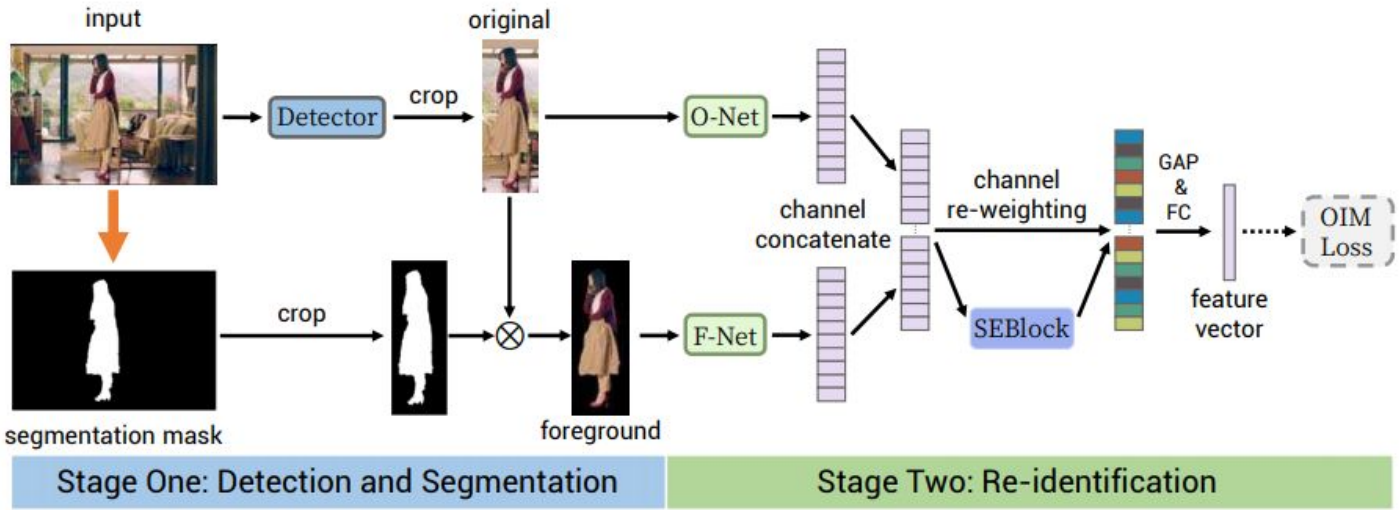
Contributions

- A deep adversarial learning architecture to jointly learn identity featrues and domain invariable features.
- Filtering low-level features and remaining high-level semantic features.
- A sketch Re-ID dataset containing 200 persons, in which each person has one sketch and two photos.

Pang, et al.., Cross-Domain Adversarial Feature Learning for Sketch Re-identification, ACM MM 2018

Pedestrian detection + Person Re-identification

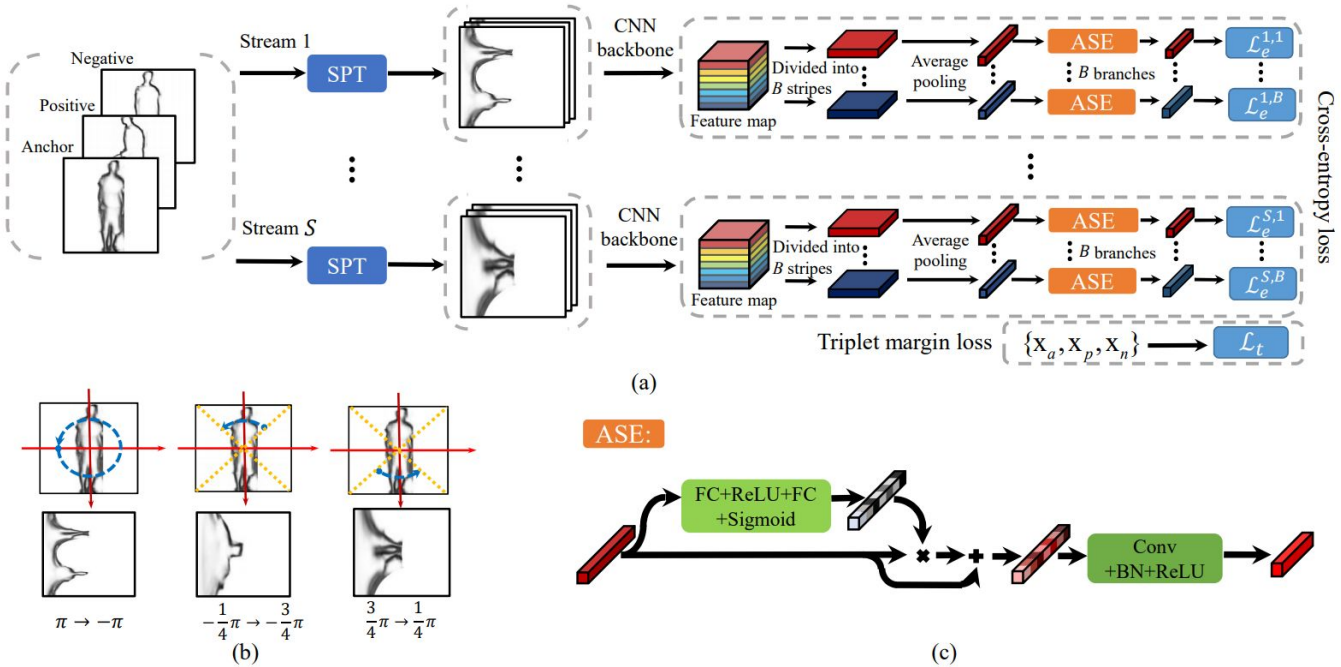
- Person search; person re-identification; person detection
- Contradict between detection and re-ID
- Background information



Chen, et al.., Person Search via A Mask-Guided Two-Stream CNN Model, ECCV 2018

Code (only dataset?)

Person re-identification with clothing change

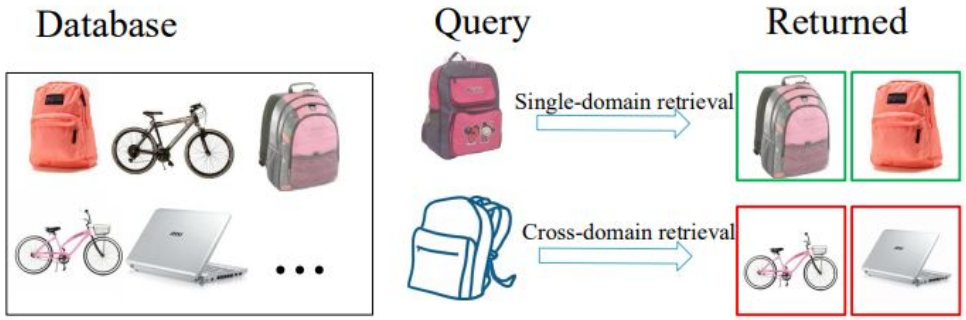


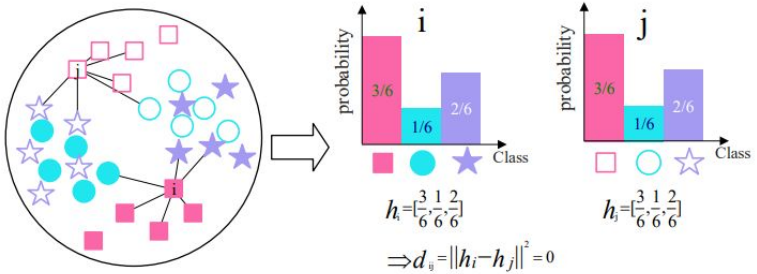
- Contributions:
  - Find **contour sketches** are much more effective for person re-id under **moderate clothing change.**
  - Introduce **SPT** to select the relatively invariant and discriminative contour patterns and **ASE** to explore angle-specific fine-grained discriminant features.
  - PRCC: A new person re-id dataset with moderate clothing changes. (221 persons)

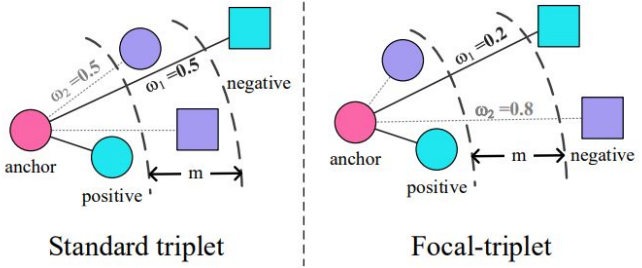
Yang, et al.., Person Re-identification by Contour Sketch under Moderate Clothing Change, TOPAMI 2019

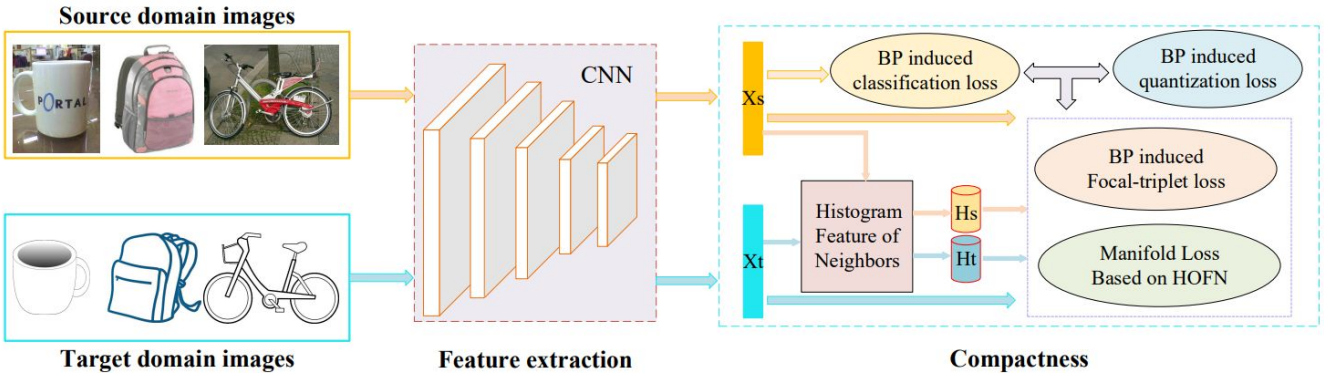
Code (matlab)

Bayesian based



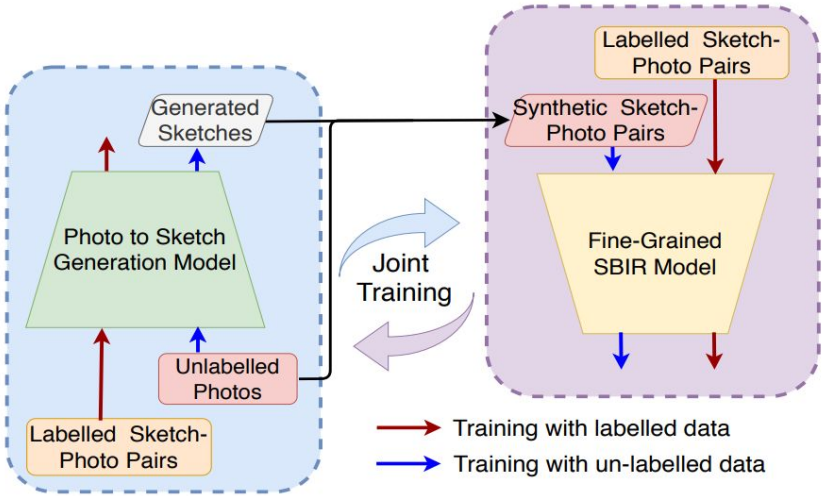






Huang, et al.., Probability Weighted Compact Feature for Domain Adaptive Retrieval, CVPR 2020 (Poster)

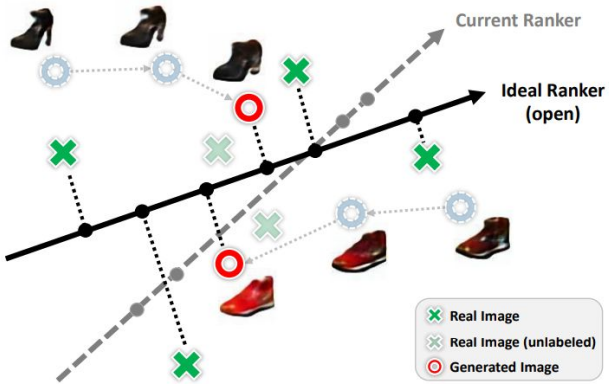
More photos are all you need



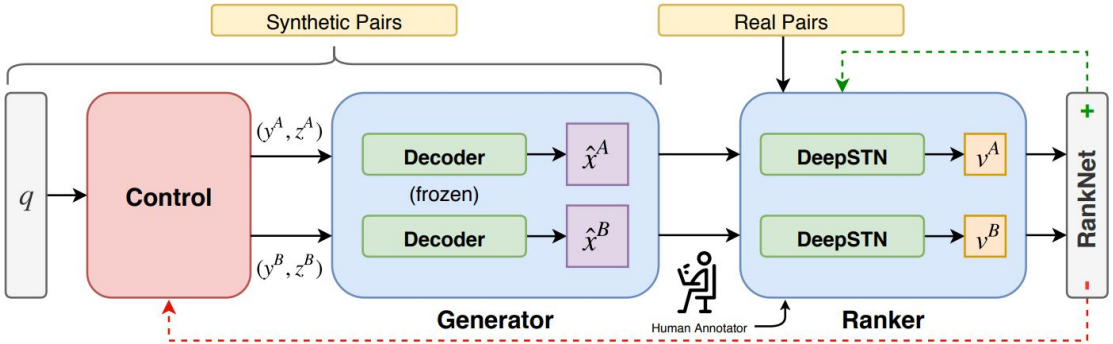
Bhunia, et al.., More photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval. CVPR 2021

Relative Attributes

ATTIC



- Contributions
  - A new approach to active learning in which pairs of samples are newly created so as to best **augment a ranker's training data**. Active generation focuses attention on novel training images that rapidly improve generalization - even after all available real images and their labels are exhausted.



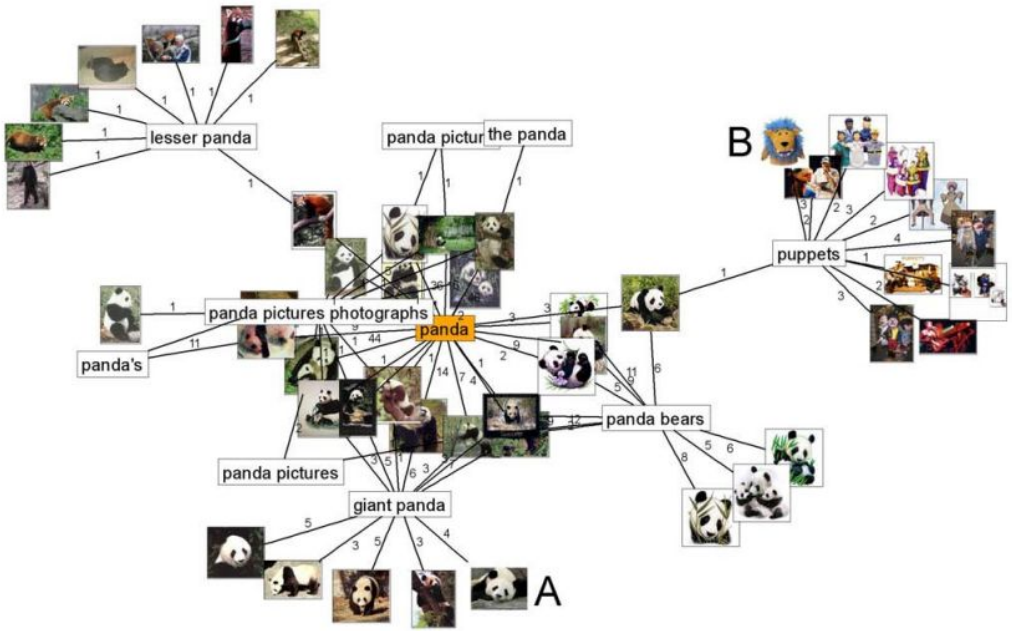
Yu, et al..,Thinking outside the pool: Active training image creation for relative attributes, CVPR 2019

Click based search

2

Random walks on Click Graph

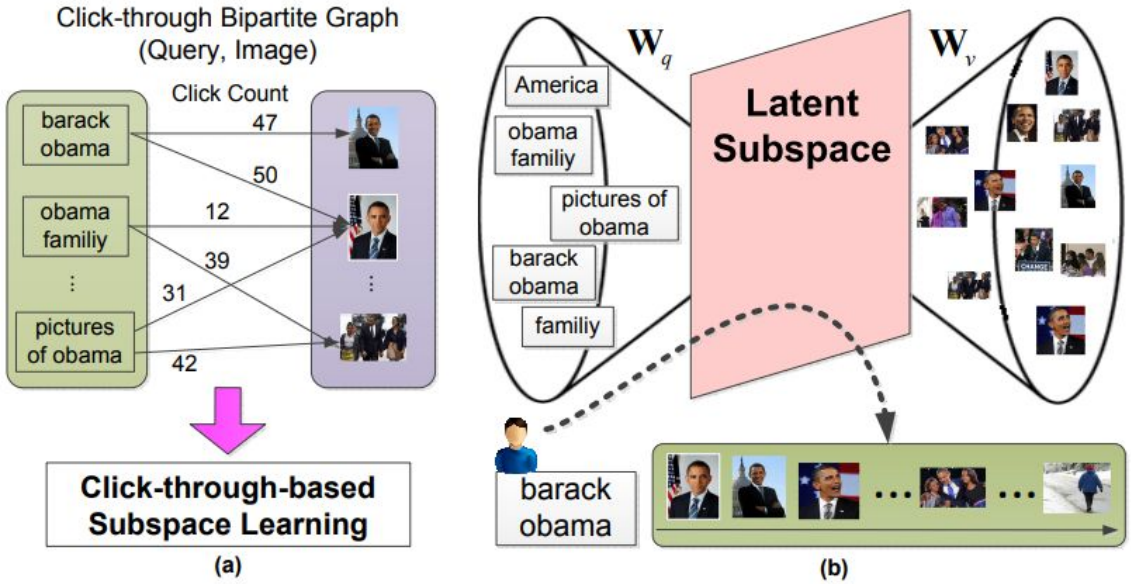
- Sparsity problem (documents that are relevant but not clicked) in clicks data



Craswell, et al.., Random Walks on the Click Graph, SIGIR 2007 (507)

Common latent subspace (text to image)+ user click data

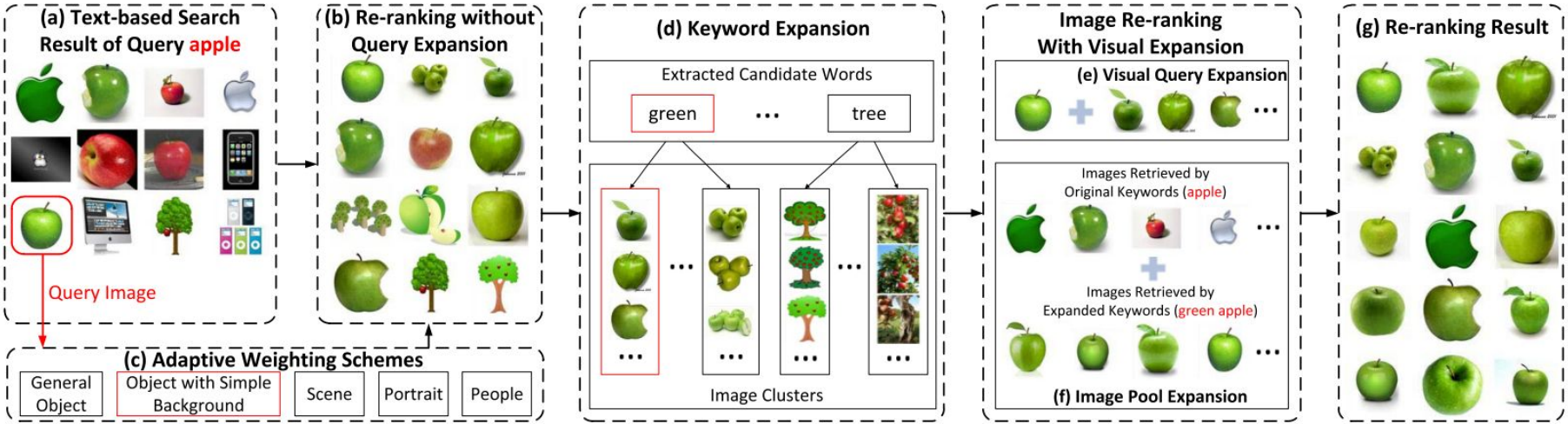
- Try to solve the ‘semantic gap’



Pan, et al.., Click-through-based Cross-view Learning for Image Search , SIGIR 2014 (87)

Pan, et al.., Click-through-based Subspace Learning for Image Search , MM 2014 (poster) (15)

One click expanse



Tang, et al.., IntentSearch: Capturing User Intention for One-Click Internet Image Search , TPAMI 2012 (148)

Video Retrieval

1

Video Moment Retrieval

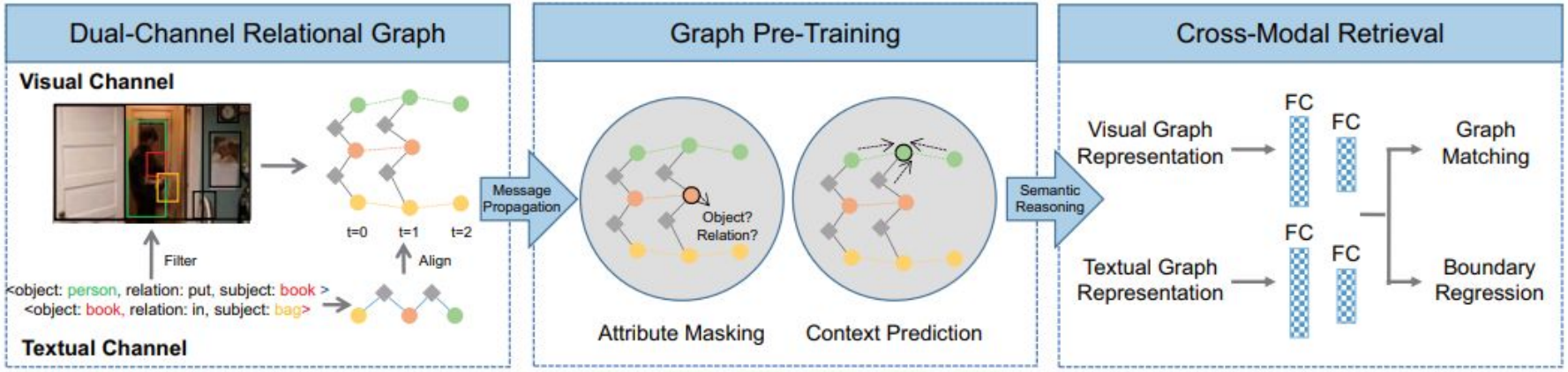
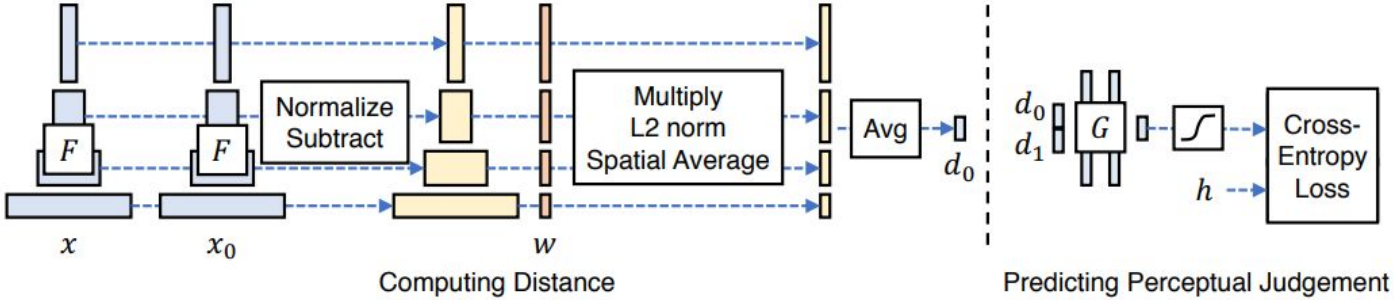


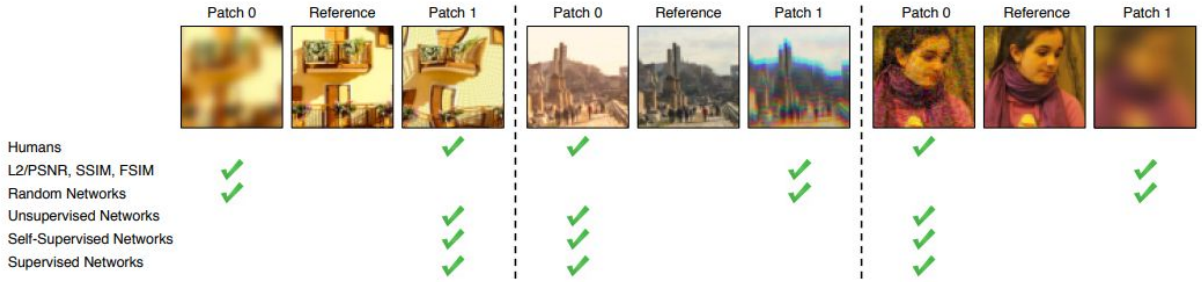
Figure 2: The graphical representation of our proposed MMRG framework. The input is an untrimmed video and its query sentence, while the output is the alignment score and location offset.

Zeng, et al., Multi-Modal Relational Graph for Cross-Modal Video Moment Retrieval, CVPR 2020

Human perceptual Metric

DataSet + Much experiment

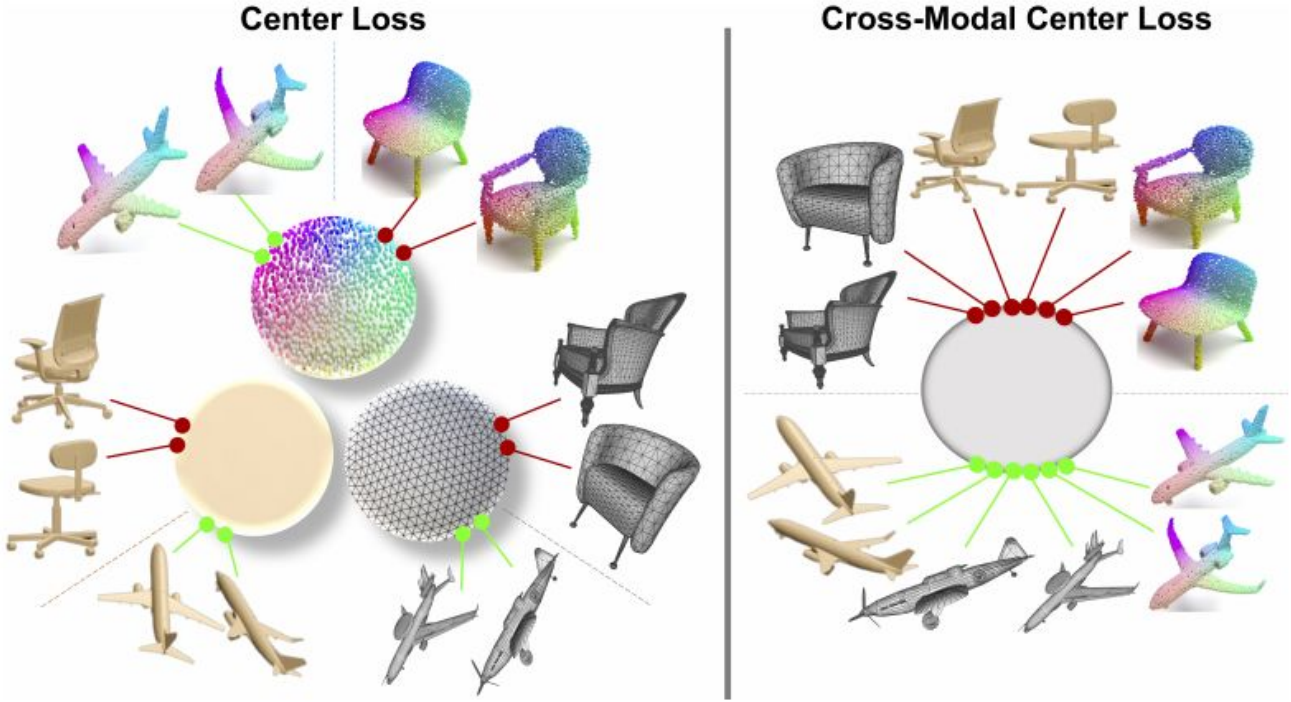




Zhang, et al.., The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, CVPR 2018 (825)

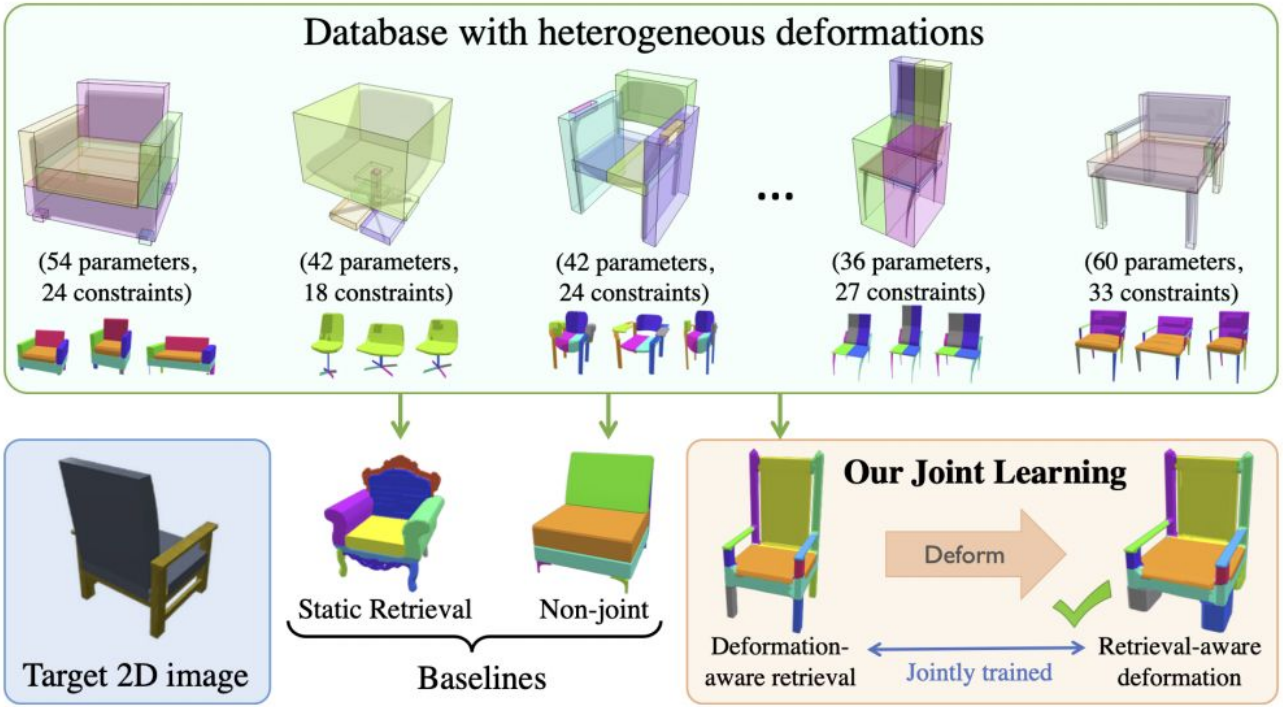
Special modal; Multi-Modal; 3D

Cross-Modal Retrieval



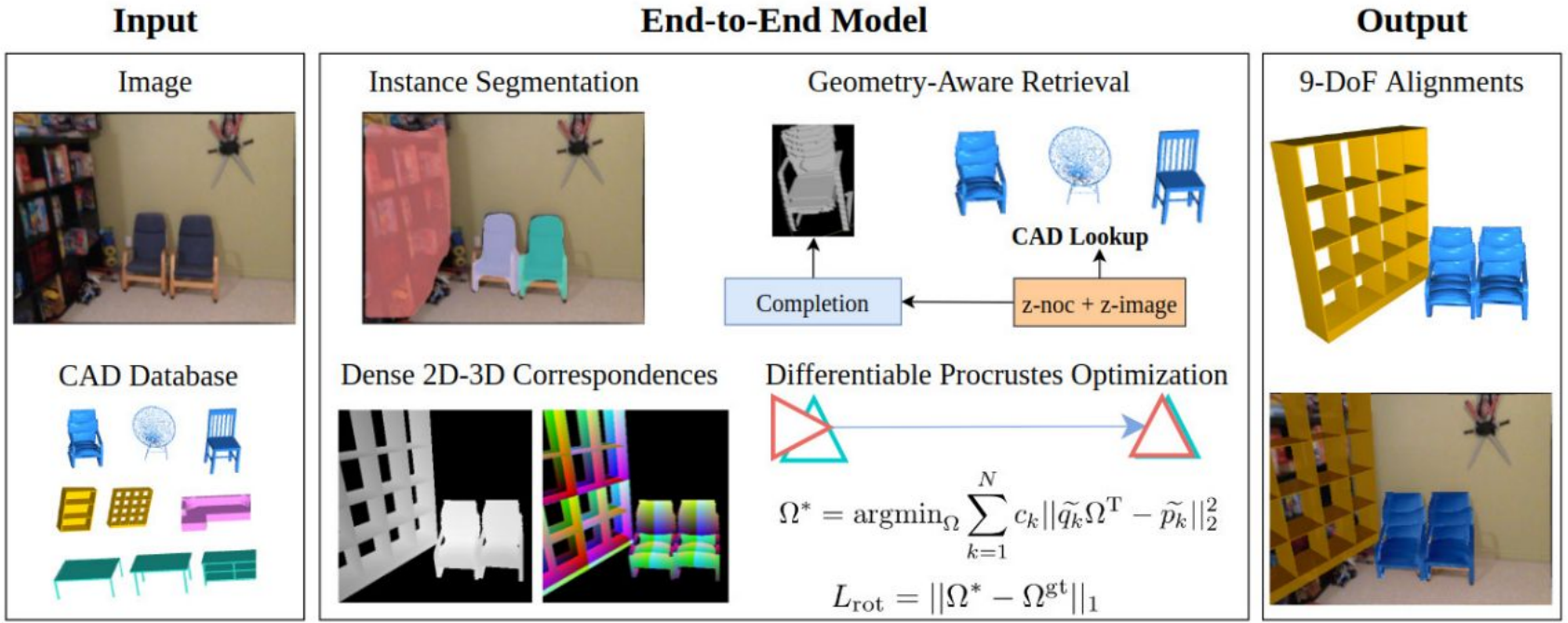
Jing, et al.., Cross-Modal Center Loss for 3D Cross-Modal Retrieval, CVPR 2021

3D shape Retrieval



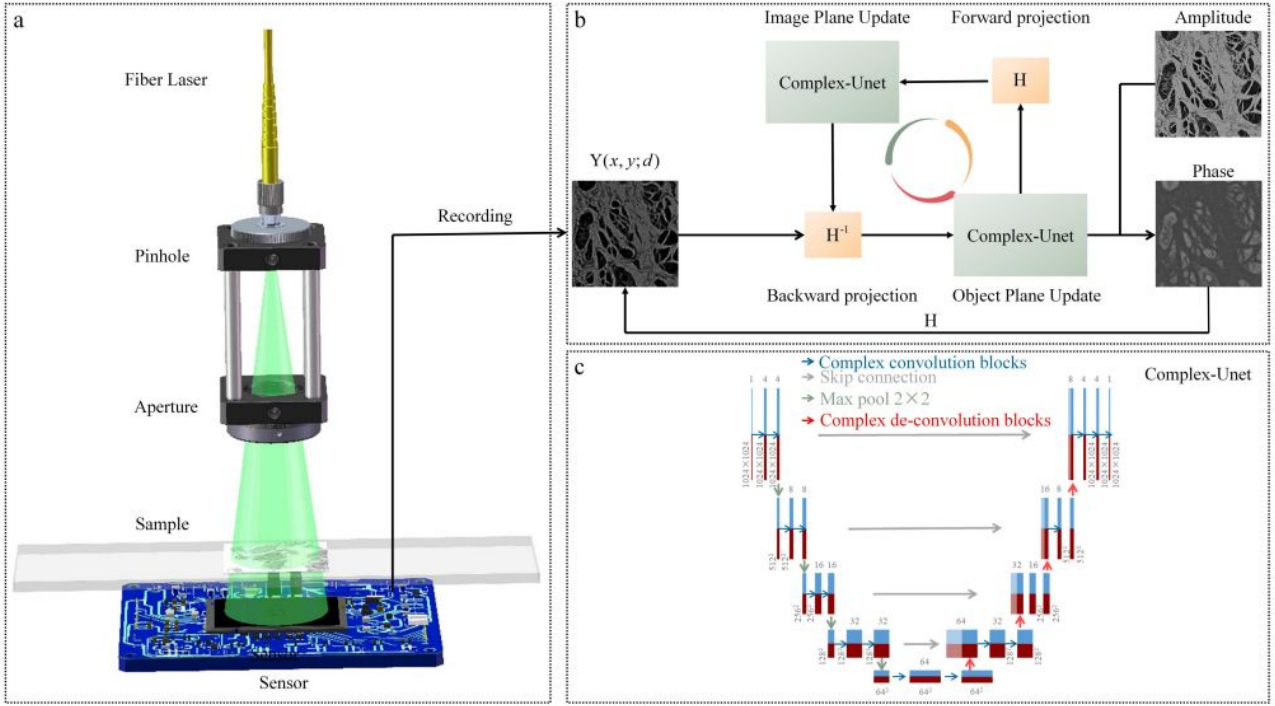
Uy, et al.., Joint Learning of 3D Shape Retrieval and Deformation, CVPR 2021

CAD Retrieval



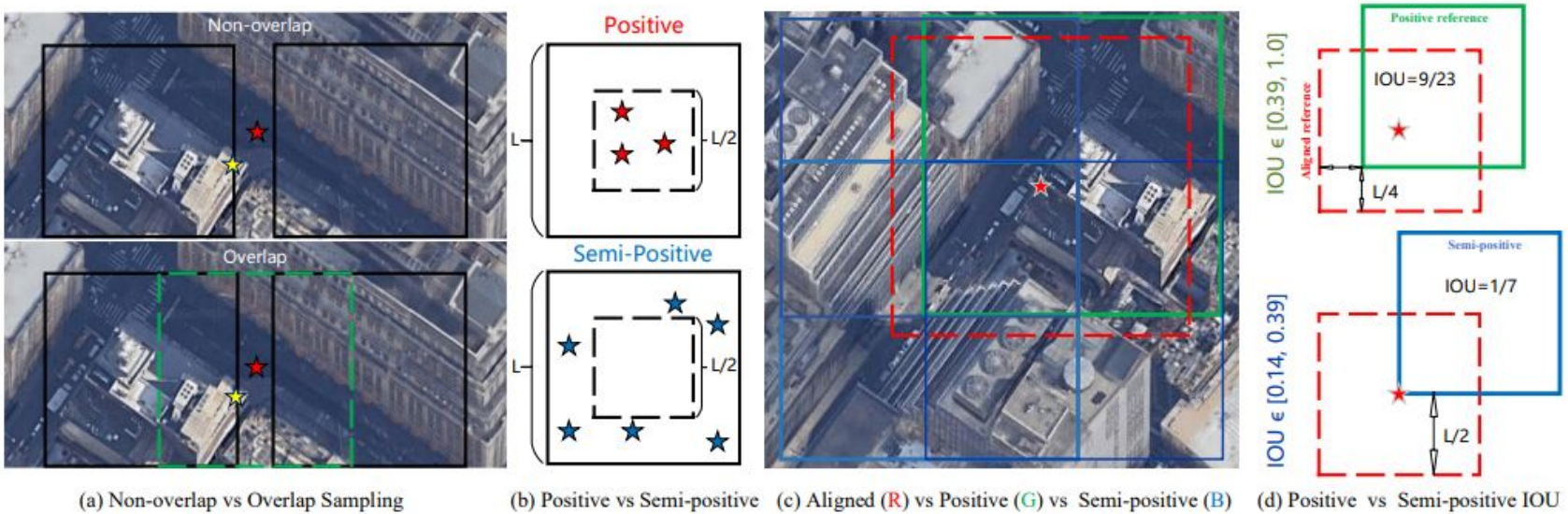
Gumeli, et al.., ROCA: Robust CAD Model Retrieval and Alignment from a Single Image, CVPR 2022

Physics Phase Retrieval



Zhang, et al...Physics-based Iteractive Projection Complex Neural Network for Phase Retrieval in Lensless Microscopy Imaging CVPR 2021

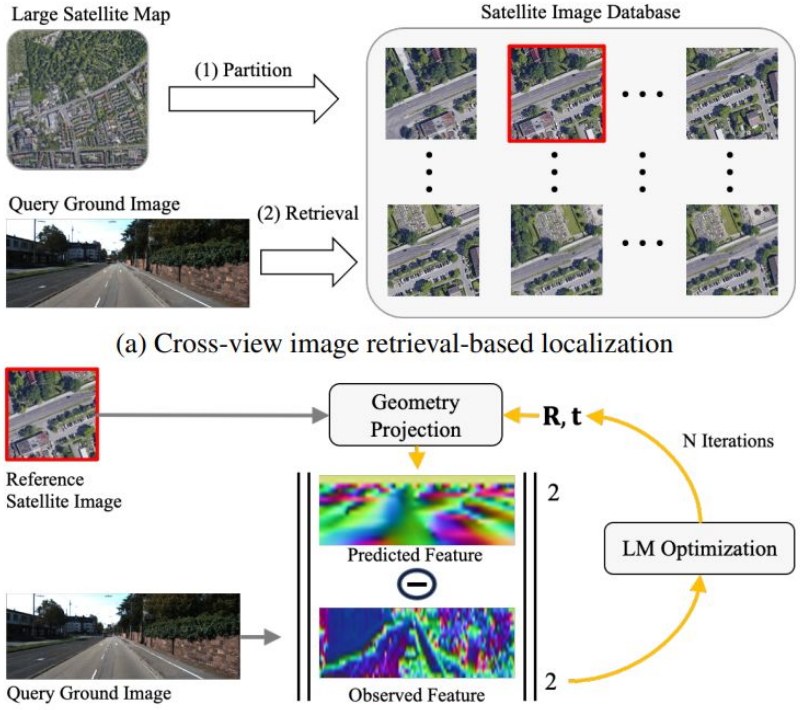
Geo-localization



(a) Non-overlap vs Overlap Sampling (b) Positive vs Semi-positive (c) Aligned (R) vs Positive (G) vs Semi-positive (B) (d) Positive vs Semi-positive IOU

Zhu, et al.,.VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval CVPR 2021

Vehicle Localization; Satellite Image

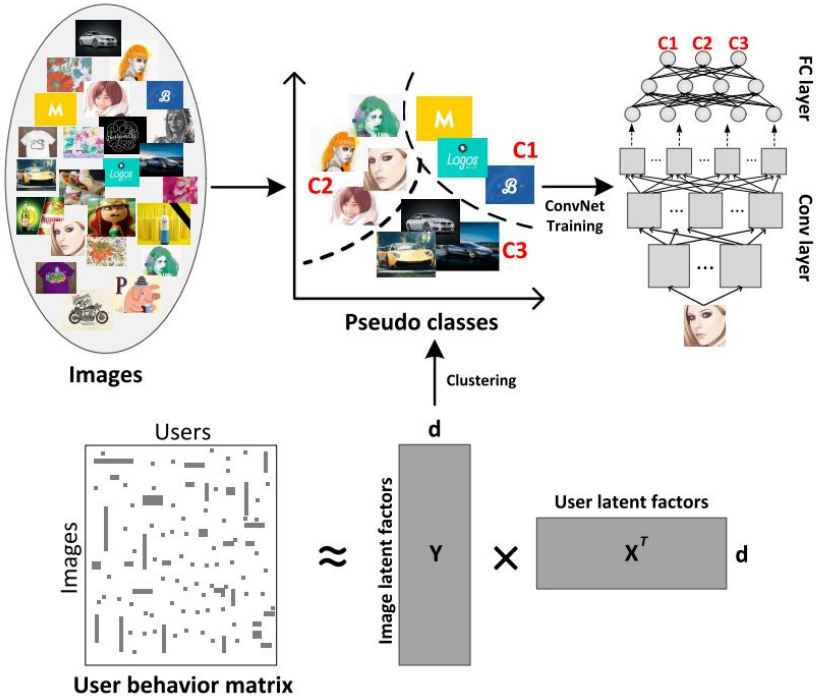


Yujiao Shi, et al..,Beyond Cross-view Image Retrieval: Highly Accurate Vehicle Localization Using Satellite Image CVPR 2022

Interactive Image Search

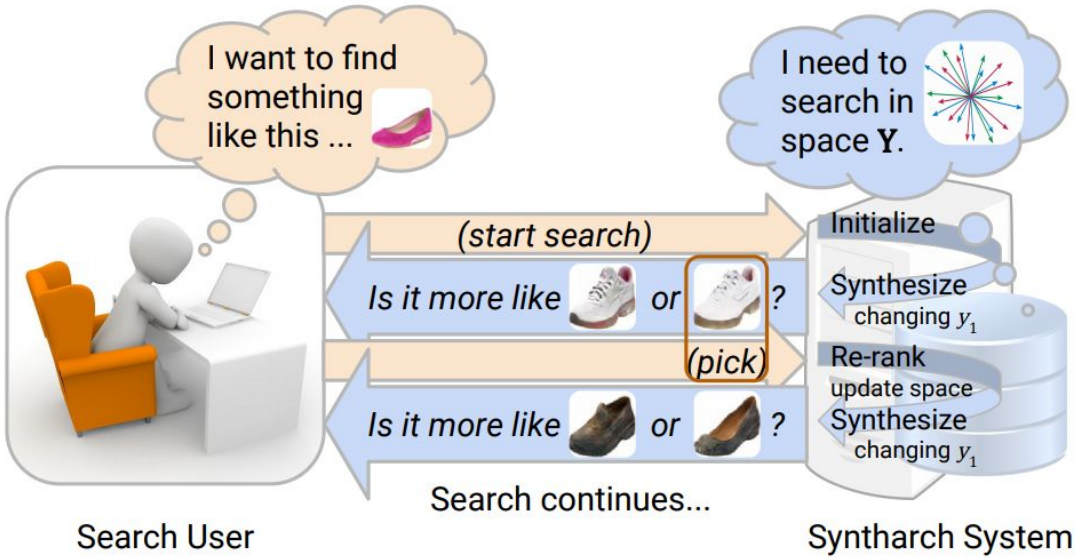
3

Image recommendation



- Contributions
  - Completely forgo the use of category labels in existing feature learning paradigms. Instead, they use user behavior data collected on social media.
  - Test on large-scale data collected from a real-world social media website.

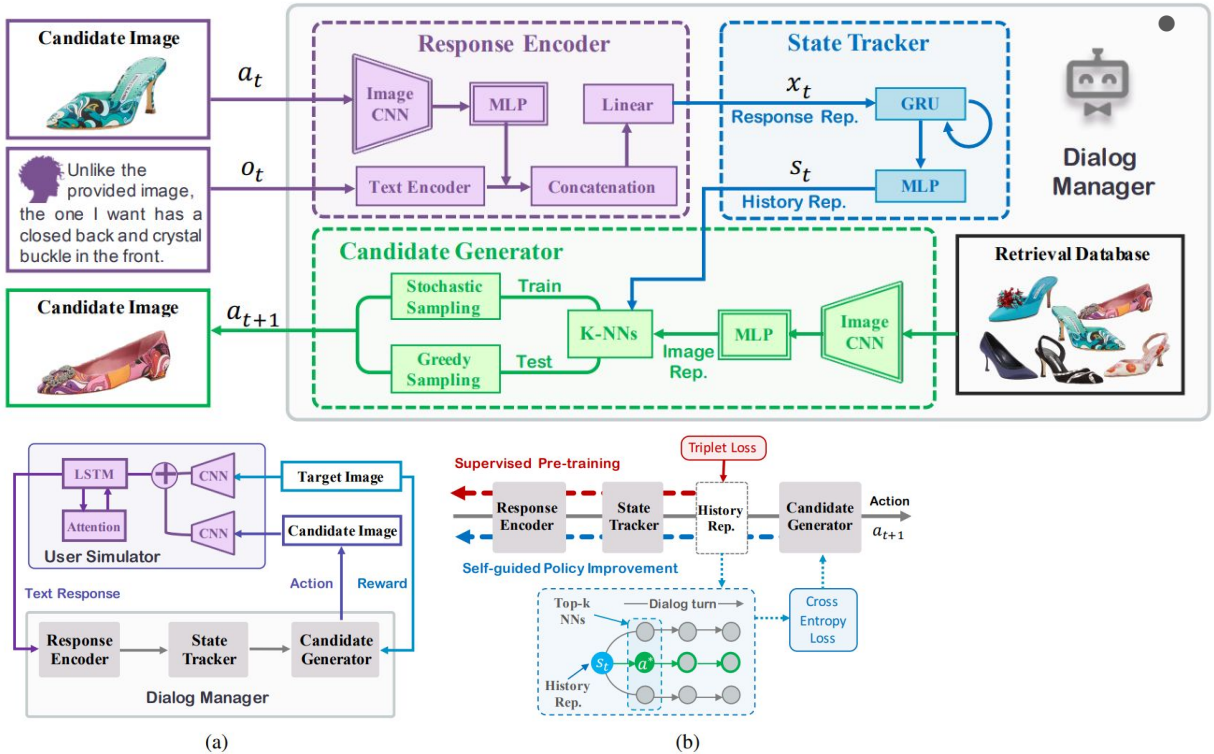
Yu, et al.., Collaborative Feature Learning from Social Media , CVPR 2015 (poster) (24)



Yu, et al.., Syntharch: Interactive Image Search with Attribute-Conditioned Synthesis, CVPR 2020 (workshop)

Code

Dialog-based Interactive Image Retrieval



Contributions

- A new vision/NLP task.
- Dialog manager architecture with triplet loss and model-based policy improvement
- A new computer vision task, relative image captioning.
- A new dataset.

Guo, et al.., Dialog-based interactive image retrieval, NIPS 2018 (32)

Murrugarra-Llerena, et al...Image retrieval with mixed initiative and multimodal feedback, BMVC 2018 (3)

Kovashka, et al.., Whittlesearch: Interactive image search with relative attribute feedback, CVPR 2012 (poster) IJCV 2015 (298)

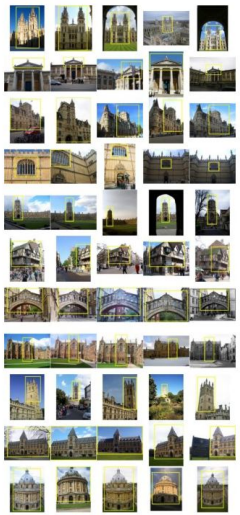
Dateset

6

Code

Oxf5k

- Detail
  - 5062 images collected from Flickr by searching for particular Oxford landmarks.
  - manually annotated; 11 different landmarks, each represented by 5 possible queries.



55 queries

Search similar images from
5062 images pool

Philbin, et al.., Object retrieval with large vocabularies and fast spatial matching, CVPR 2007

Code

Paris and Flicker1

Dataset	Number of images	Number of features
Oxford	5,062	16,334,970
Paris	6,300	20,219,488
Flickr1	99,782	277,770,833
Total	111,144	314,325,291

Table 1. The number of descriptors for each dataset.

Philbin, et al.., Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. CVPR 2008

Code

R-Oxf & R-Par

- Detail
  - **ROxford**: 4993 images + 70 queries.

Radenovic, et al.., Revisiting Oxford and Paris: Large_Scale Image Retrieval Benchmarking, CVPR 2018

Code

GLDv2

- Detail
  - **GLDv2-retrieval**: 1129 queries(379 validation and 750 testing) and 762k database images
  - **GLDv2-recognation**: 118K test (41k validation and 77k testing) and 4M training images from 203k landmarks.

Dataset name	Year	# Landmarks	# Test images	# Train images	# Index images	Annotation collection	Coverage	Stable
Oxford [41]	2007	11	55	-	5k	Manual	City	Y
Paris [42]	2008	11	55	-	6k	Manual	City	Y
Holidays [28]	2008	500	500	-	1.5k	Manual	Worldwide	Y
European Cities 50k [5]	2010	20	100	-	50k	Manual	Continent	Y
Geotagged StreetView [32]	2010	-	200	-	17k	StreetView	City	Y
Rome 16k [1]	2010	69	1k	-	15k	GeoTag + SfM	City	Y
San Francisco [14]	2011	-	80	-	**1.7M**	StreetView	City	Y
Landmarks-PointCloud [35]	2012	1k	10k	-	205k	Flickr label + SfM	Worldwide	Y
24/7 Tokyo [55]	2015	125	315	-	1k	Smartphone + Manual	City	Y
Paris500k [60]	2015	13k	3k	-	501k	Manual	City	N
Landmark URLs [7, 22]	2016	586	-	140k	-	Text query + Feature matching	Worldwide	N
Flickr-SfM [44]	2016	713	-	120k	-	Text query + SfM	Worldwide	Y
Google Landmarks [39]	2017	30k	**118k**	1.2M	1.1M	GPS + semi-automatic	Worldwide	N
Revisited Oxford [43]	2018	11	70	-	5k + 1M	Manual + semi-automatic	Worldwide	Y
Revisited Paris [43]	2018	11	70	-	6k + 1M	Manual + semi-automatic	Worldwide	Y
Google Landmarks Dataset v2	2019	**200k**	**118k**	**4.1M**	762k	Crowsourced + semi-automatic	Worldwide	Y

Table 1: Comparison of our dataset against existing landmark recognition/retrieval datasets. "Stable" denotes if the dataset can be retained indefinitely. Our Google Landmarks Dataset v2 is larger than all existing datasets in terms of total number of images and landmarks, besides being stable.

Weyand, et al.., Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, CVPR 2020

Code

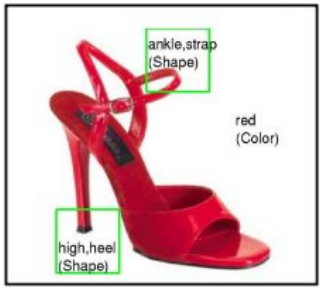
Attribute Discovery Dataset

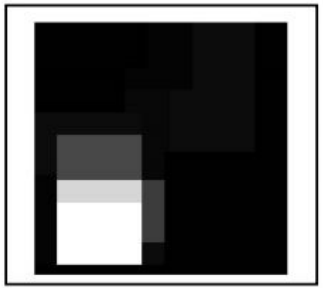
- Detail
  - 4 broad shopping categories (bags, earrings, ties, and shoes).
  - 37795 images: 9145 handbags; 14765 shoes; 9235 earrings; 4650 ties.
  - Collected from like.com
  - images and associated textual descriptions

Dazzle after dark with Judith Leiber's decadent oversized crystal-embellished silver-tone clutch. Carry this fabulous extra to add high-octane glamour to an LBD and teetering heels. Shown here with an Emilio Pucci dress and Givenchy shoes.

The 12K pink and green gold leaves gently cascade down on these delicate beaded 10K gold earrings.

Rock and roll in th
Signature. The sm
leather upper with
atop a 1 inch plat
ankle with a gold
heel. Sizzle in the





Berg, et al.., Automatic Attribute Discovery and Characterization from Noisy Web Images, ECCV 2010

Code

Relative captioning dataset

- Detail
  - Shoes dataset. 14658 images
  - Relative_captions_shoes.json 10751 captions, with one caption per pair of images.
  - Captions_shoes.json: captions on 3600 shoe images

(b)

Guo, et al.., Dialog-based interactive image retrieval, NIPS 2018 (32)

UT-Zap50K

Yu, et al.., Fine-grained visual comparisons with local learning, CVPR 2014 (247)

Yu, et al..,Semantic jitter: Dense supervision for visual comparisons via synthetic images, ICCV 2017 (56)

DeepFashion

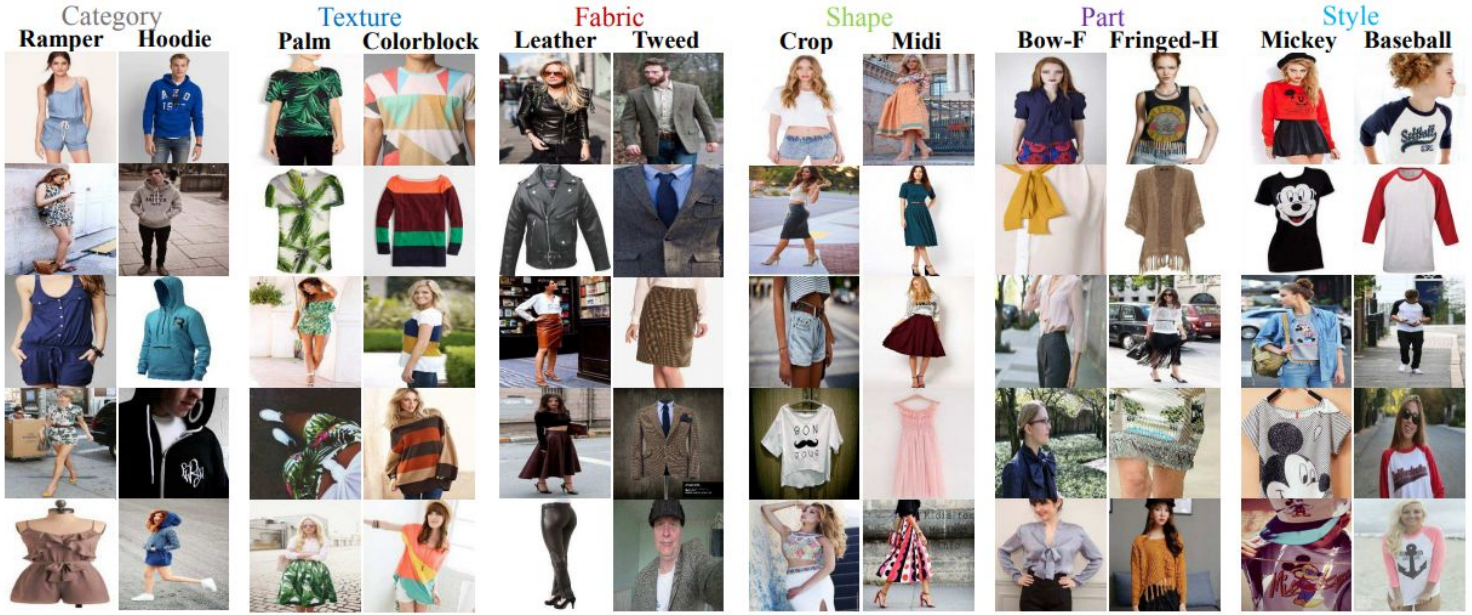


Figure 2. Example images of different categories and attributes in DeepFashion. The attributes form five groups: texture, fabric, shape, part, and style.

Liu, et al.., DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations, CVPR 2016 (696)

DARN

Huang, et al..,Cross-domain Image Retrieval with a Dual Attribute-aware Ranking Network, ICCV 2015 (202)

To read