Deep Image Retrieval – COMP4423 Computer Vision

Xiaoyong Wei (魏驍勇)
x1wei@polyu.edu.hk

Department of Computing
電子計算學系

The Hong Kong Polytechnic University
香港理工大學

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Outline

>Deep image retrieval

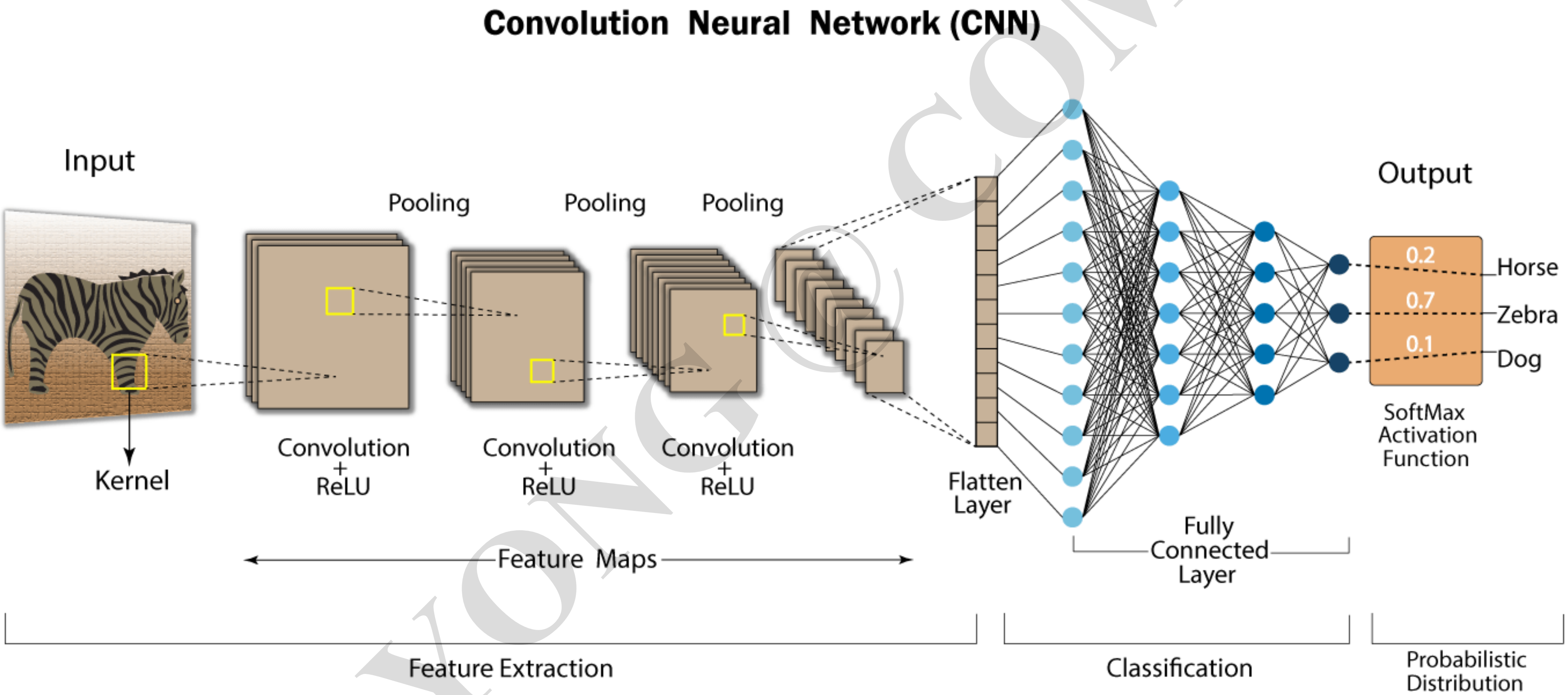
>Feature aggregation/embedding/fusion

>Fine tuning (Siamese/Triplet networks)

Deep Learning is cool. It's in fact a game changer not only for classification, but also a wide range of Computer Vision tasks.

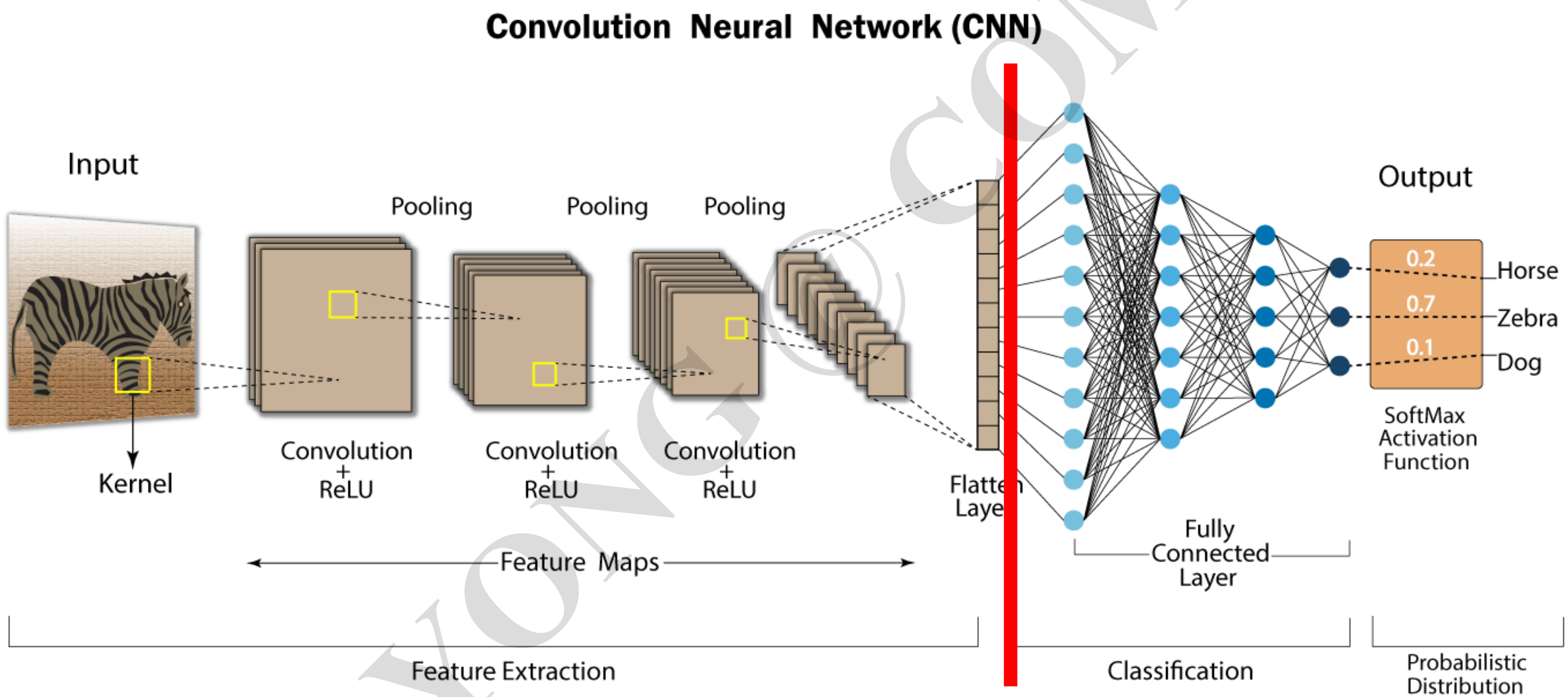
Let's see how it helps the image retrieval

Convolutional Networks



https://discuss.boardinfinity.com/t/what-do-you-mean-by-convolutional-neural-network/8533

Convolutional Networks

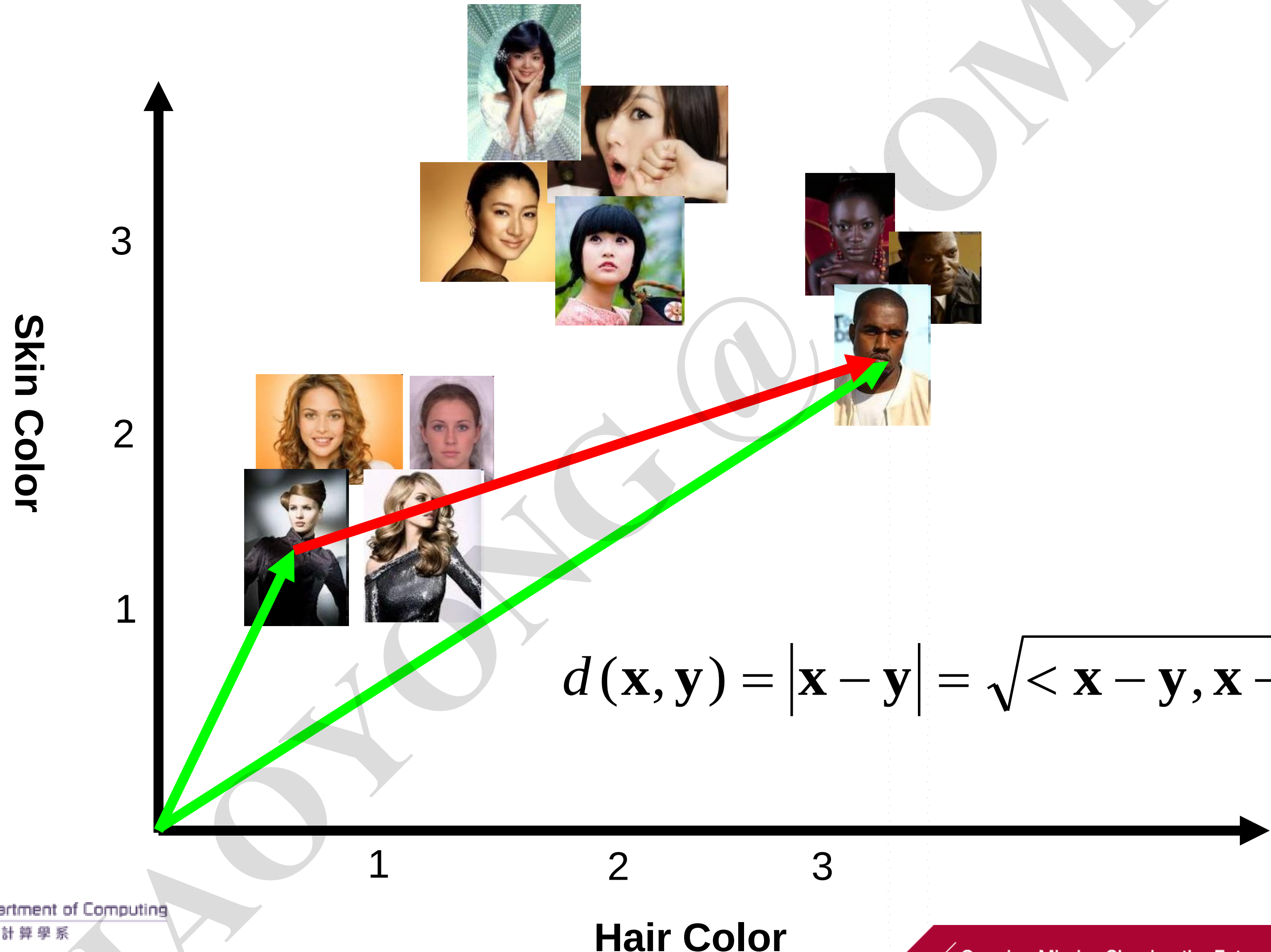


Up to here, the images are converted into feature vectors (represented in the feature space).

https://discuss.boardinfinity.com/t/what-do-you-mean-by-convolutional-neural-network/8533

As mentioned early, as long as the images are represented in the feature space, the search can be conducted by ranking images using similarity/distance metrics.

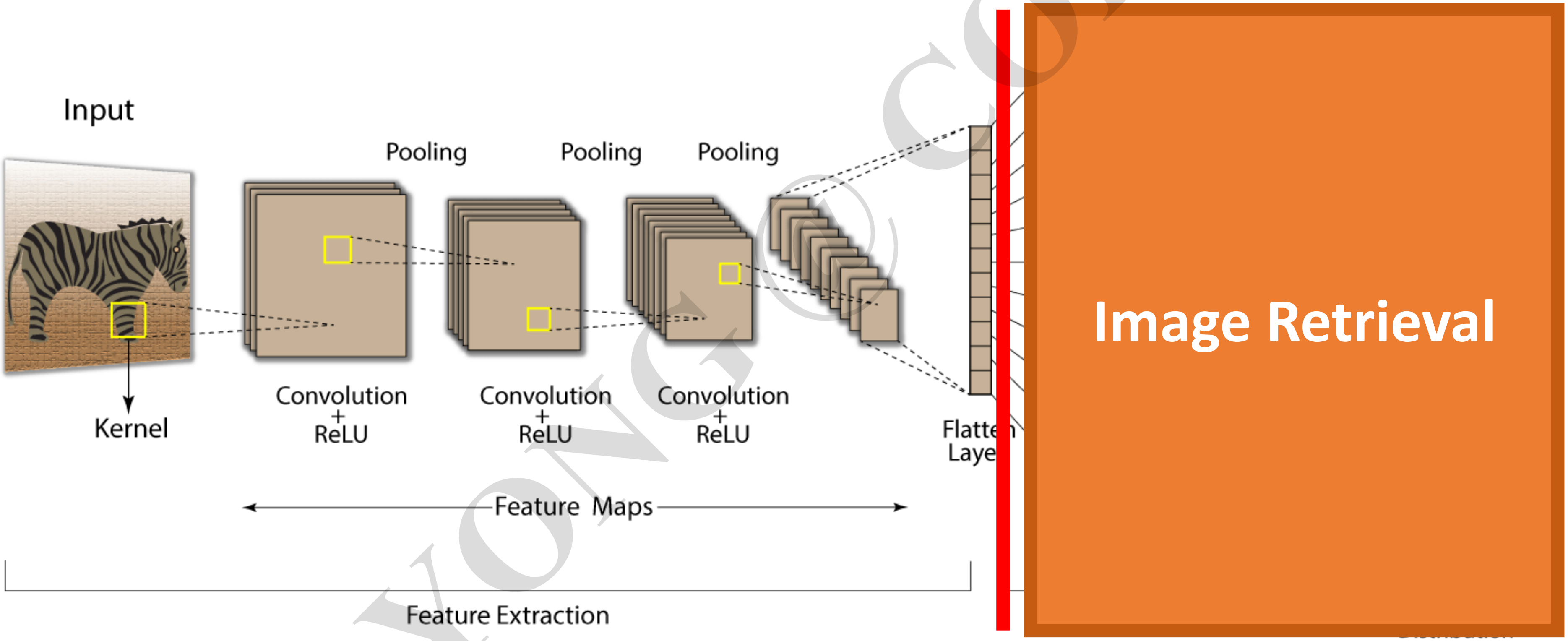
Metrics – Euclidian Distance



$$d(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}| = \sqrt{<\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}>}$$

The only difference is that the feature space is now spanned by **deep** feature vectors.

Deep Image Retrieval



Up to here, the images are converted into feature vectors (represented in the feature space).

https://discuss.boardinfinity.com/t/what-do-you-mean-by-convolutional-neural-network/8533

What's the best way of using deep features?

Can we construct better features instead of using the raw feature maps?

Feature Aggregation

In feature maps the spatial dimensions of the original images are “preserved”. We can thus **summarize** the features over the spatial dimensions for better representations of regions. This can be done by using different types of pooling algorithms.

5	3	1	2
1	2	3	2
4	2	2	5
3	6	1	1

→

2.75	2
3.75	2.25

Sum/average Pooling

5	3	1	2
1	2	3	2
4	2	2	5
3	6	1	1

→

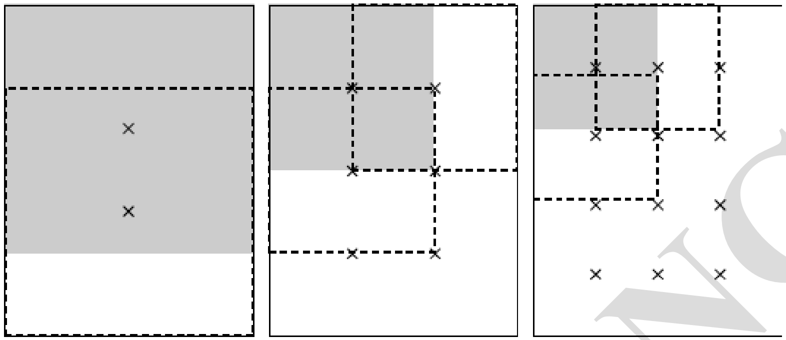
5	3
6	5

Max Pooling

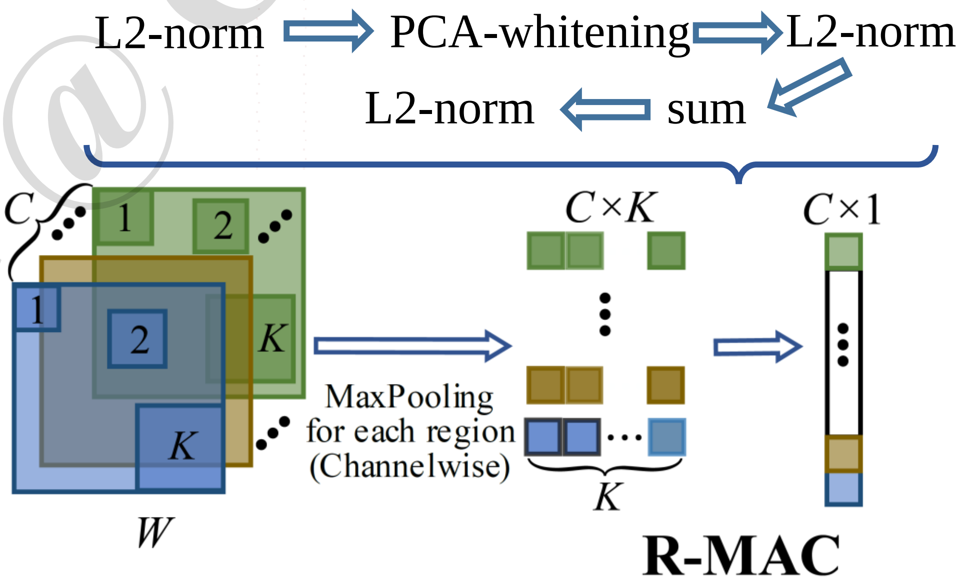
Chen W, Liu Y, Wang W, et al. Deep image retrieval: A survey[J]. arXiv preprint arXiv:2101.11282, 2021.

Single Forward-Forward Pass

R-MAC derives a compact image representation from the convolutional layers to encode multiple image regions

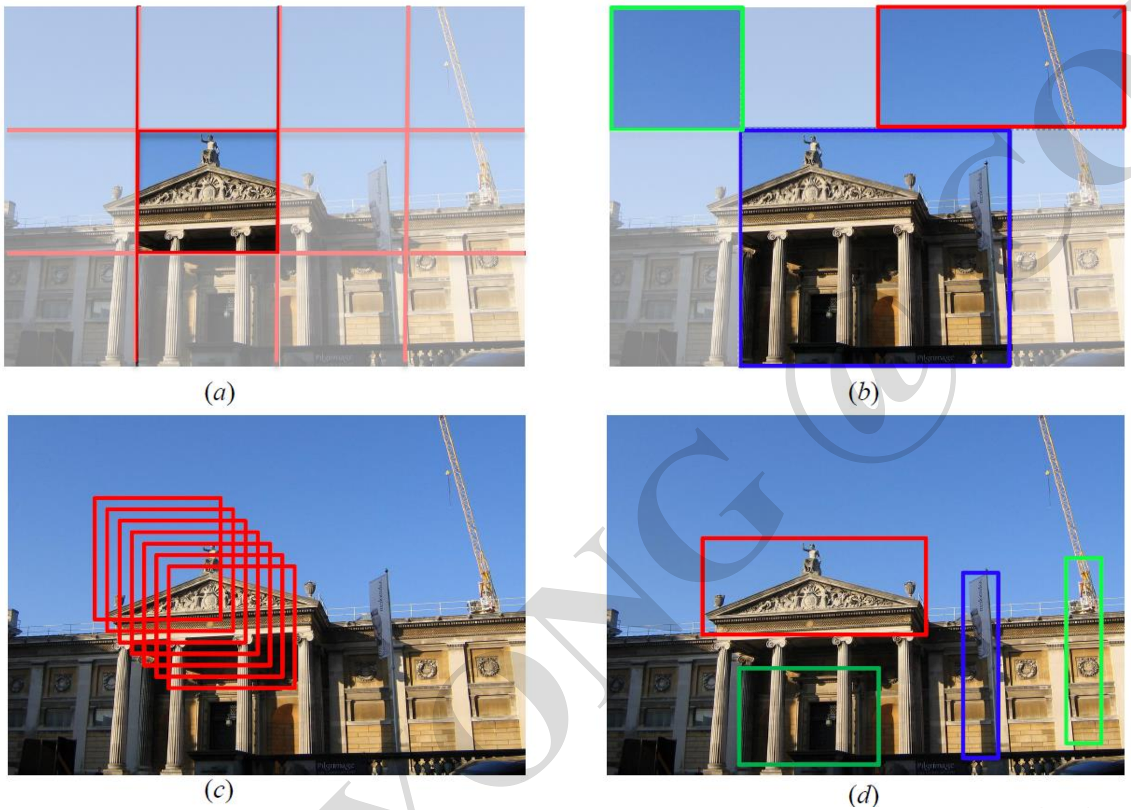


The regions are sampled uniformly with overlaps between consecutive regions



Chen W, Liu Y, Wang W, et al. Deep image retrieval: A survey[J]. arXiv preprint arXiv:2101.11282, 2021.
Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations[J]. arXiv preprint arXiv:1511.05879, 2015.

Multiple Feed-Forward Pass



Advantage: higher retrieval accuracy

Disadvantage: time-consuming

(a) Rigid grid; (b)Spatial pyramid modeling (SPM); (c) Dense patch sampling;(d) Region proposals (RPs) from region proposal networks.

Feature Embedding

In addition to the direct pooling or regional pooling, we can use embedding to convert the feature maps into compact features. Representative methods include: **BoW, VLAD,** and **FV** .

VLAD generates K visual word centroids, assigns each feature $\vec{x}_t$ to its nearest visual centroid $\vec{c}_k$, and aggregates the difference $(\vec{x}_t, \vec{c}_k)$ as

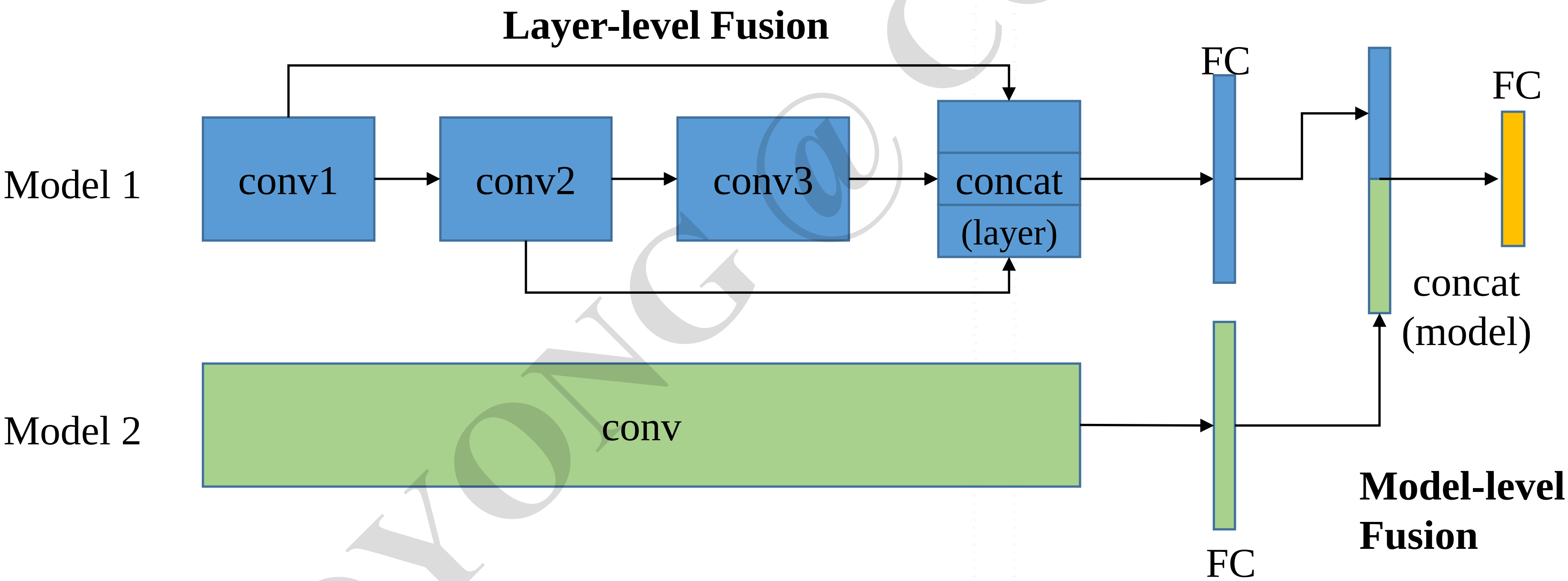
$$g(\vec{c}_k) = \frac{1}{T}\sum_{t=1}^{T} \phi(\vec{x}_t, \vec{c}_k)(\vec{x}_t - \vec{c}_k)$$

$$\phi(\vec{x}_t, \vec{c}_k) = \begin{cases} 1, if\ \vec{c}_k\ is\ the\ nearest\ codeword\ for\ \vec{x}_t \\ 0, therwise \end{cases}$$

The VLAD representation is stacked with the residuals to all centroids, with dimension $(D \times K)$:

$$G_{VLAD}(\vec{x}) = [\ldots, g(\vec{c}_k)^{\intercal}, \ldots]^{\intercal}$$

Feature Fusion



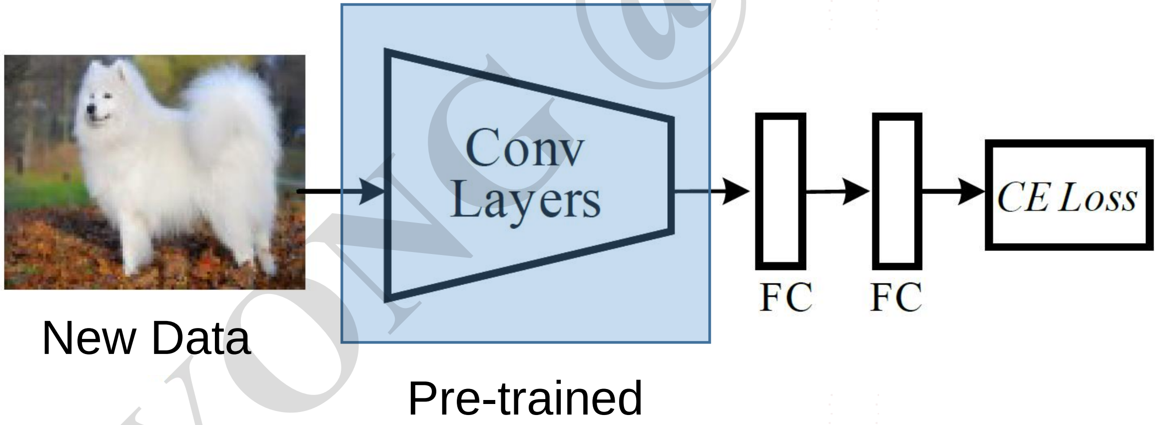
Chen W, Liu Y, Wang W, et al. Deep image retrieval: A survey[J]. arXiv preprint arXiv:2101.11282, 2021.

These methods are in fact called **off-the-shell** methods, because they don't change the parameters (weights) of the original CNNs.

By contrast, there are **fine-tuned methods**, in which we can update the parameters (weights) for better performance (to address the **domain shift**).

Classification-based Tuning

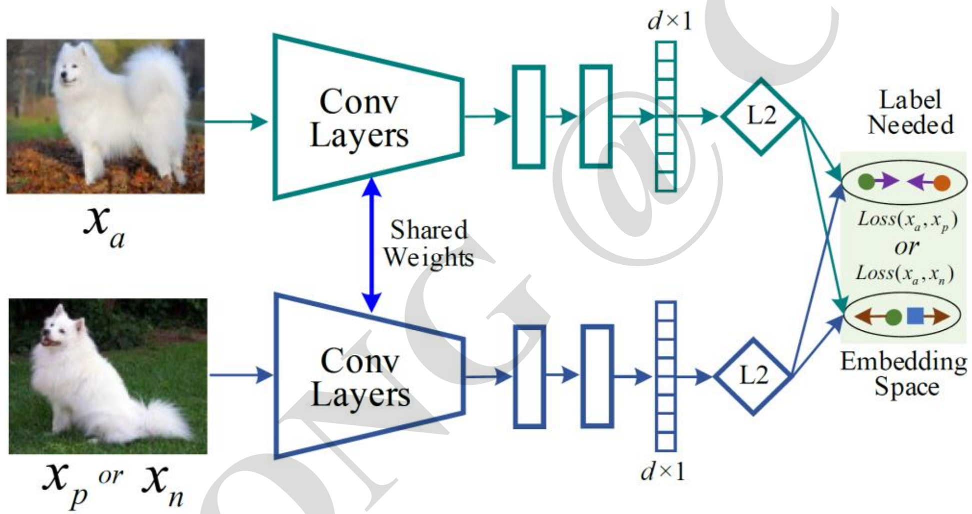
Retrain the pre-trained DCNN (AlexNet, VGG, GoogLeNet, or ResNet) when labels on the new datasets are available so as to improve the model-level adaptability on the new datasets.



Chen W, Liu Y, Wang W, et al. Deep image retrieval: A survey[J]. arXiv preprint arXiv:2101.11282, 2021.

Verification-based Tuning

1) A pair-wise constraint (e.g., Siamese network)

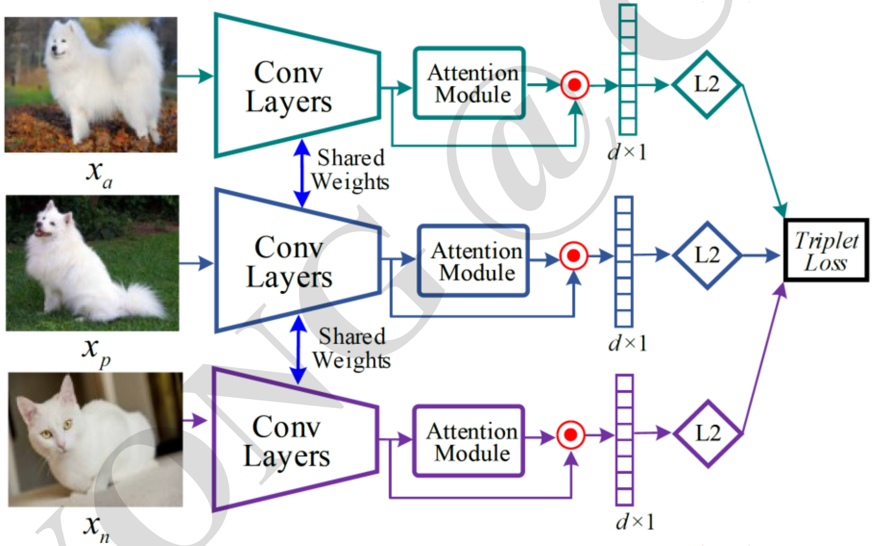


$$L_{Siam}(x_i, x_j) = \frac{1}{2}S(x_i, x_j)D(x_i, x_j) + \frac{1}{2}(1 - S(x_i, x_j))\max(0,\ m - D(x_i, x_j))$$

$$D(x_i, x_j) = ||f(x_i; \boldsymbol{\theta}) - f(x_j; \boldsymbol{\theta})||_2^2 \qquad S(x_i, x_j) \in \{0, 1\}$$

Verification-based Tuning

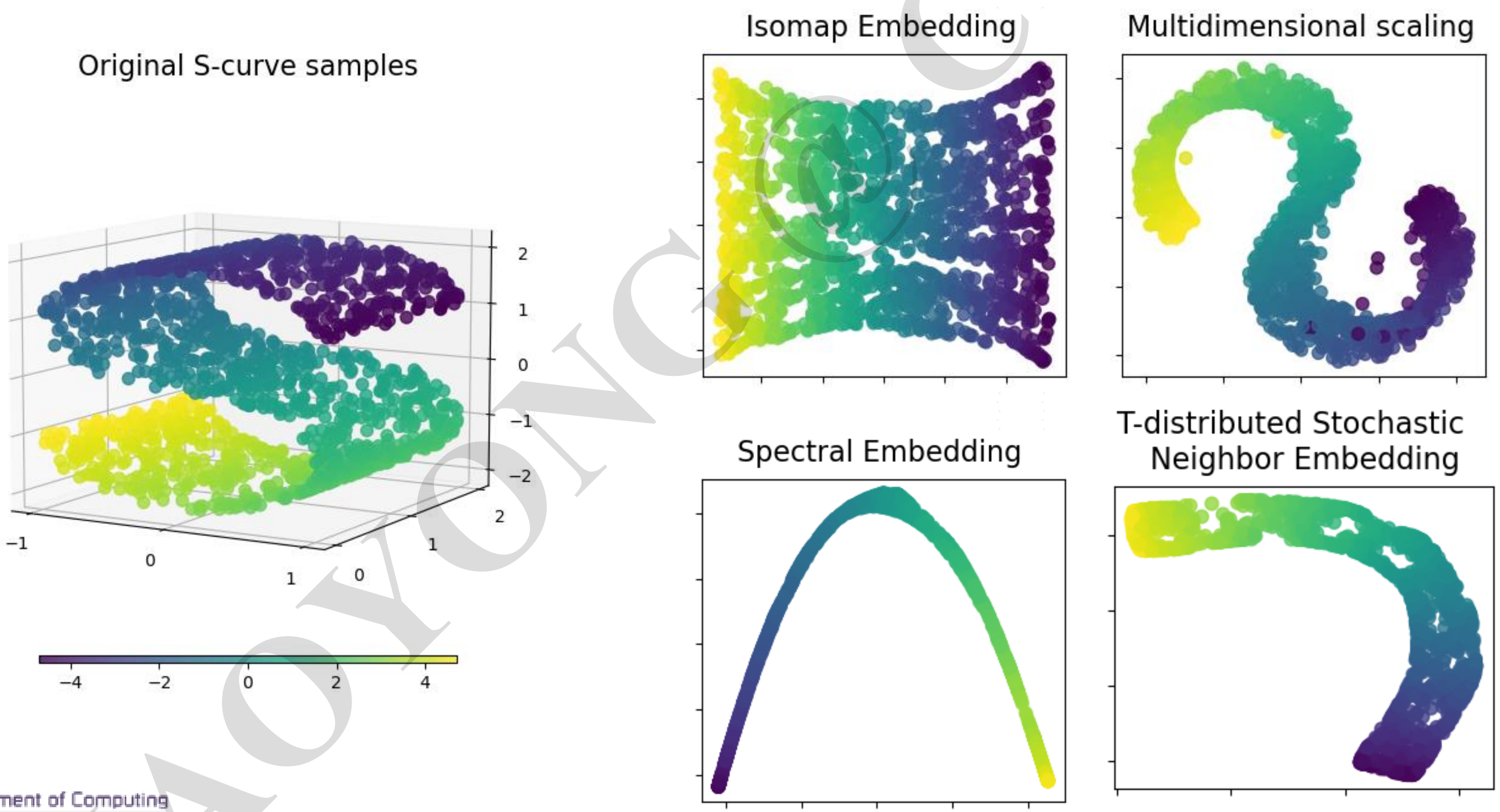
2) A triplet constraint (e.g., triplet networks)



$$L_{Triplet}(x_a, x_p, x_n) = \max(0, m + D(x_a, x_p) - D(x_a, x_n)))$$

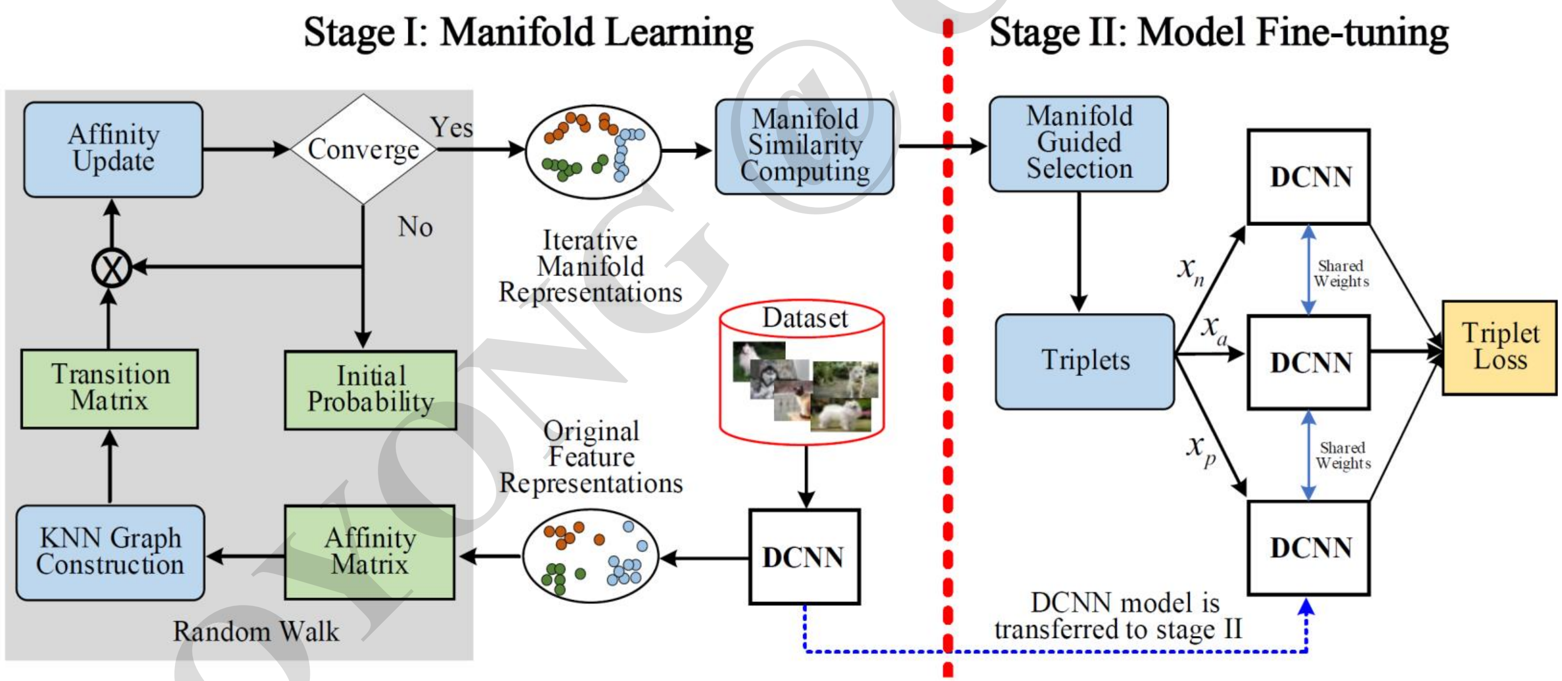
Unsupervised Tuning

Manifold learning is a method for non-linear dimensionality reduction, which learns the intrinsic correlation of data in a high dimensional space and represent them in a low dimensional space (with the correlation preserved).

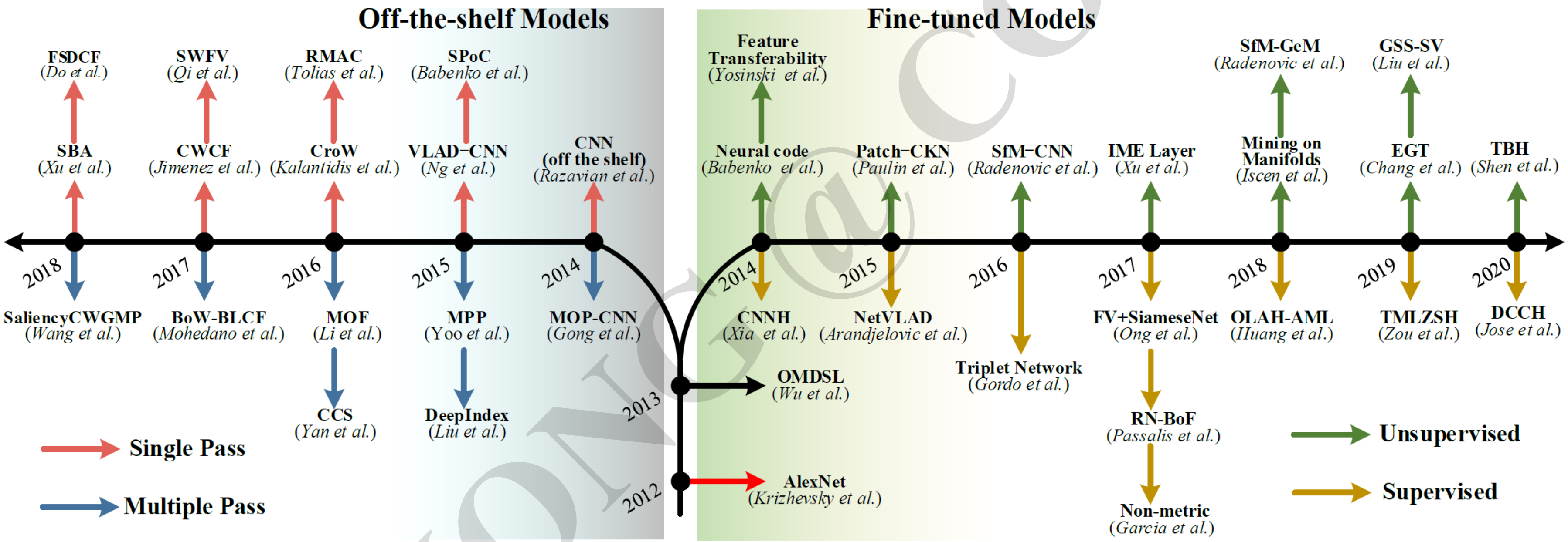


Unsupervised Tuning

Manifold learning is used as a way to guide the sampling of positive and negative pairs.



Deep Image Retrieval



Chen W, Liu Y, Wang W, et al. Deep image retrieval: A survey[J]. arXiv preprint arXiv:2101.11282, 2021.

Thank you!