MIPR’21

2020 TCMC Impact Award

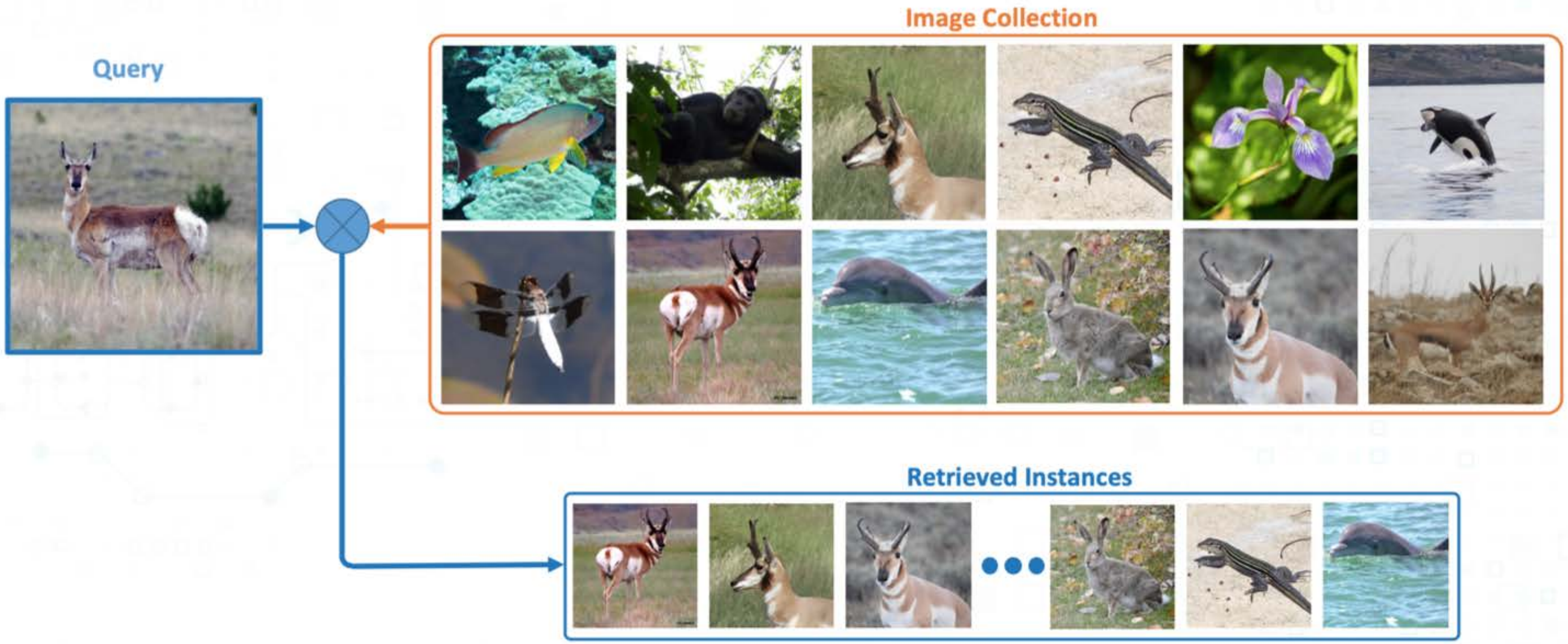
Bridging Gap between Image Pixels and Semantics via Supervision

C.-C. Jay Kuo

University of Southern California

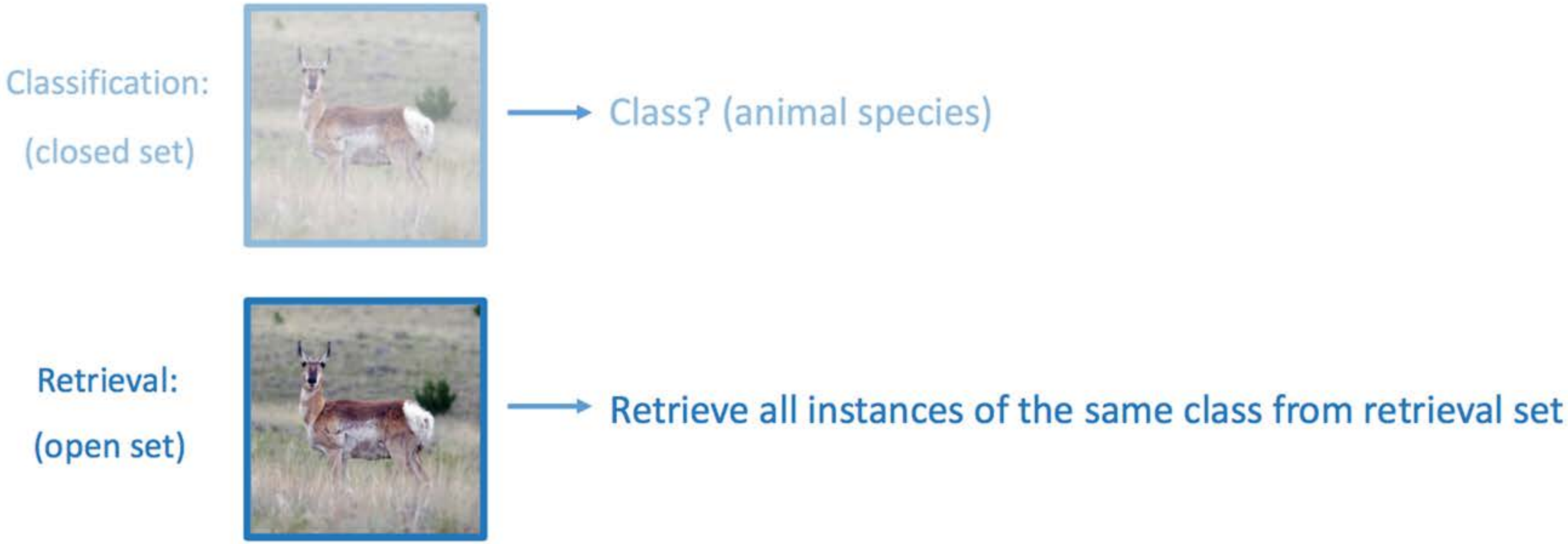
Content-Based Image Retrieval (CBIR)

- **What**: Given a query, find relevant images from the database
- **How**: Compare similarity of representative features



Comparison between Detection and Retrieval

- Detection vs. Classification
  - Open set vs. Closed set



Challenges in Semantic Annotation
(A slide taken from 20 years ago)

This image can be annotated by:

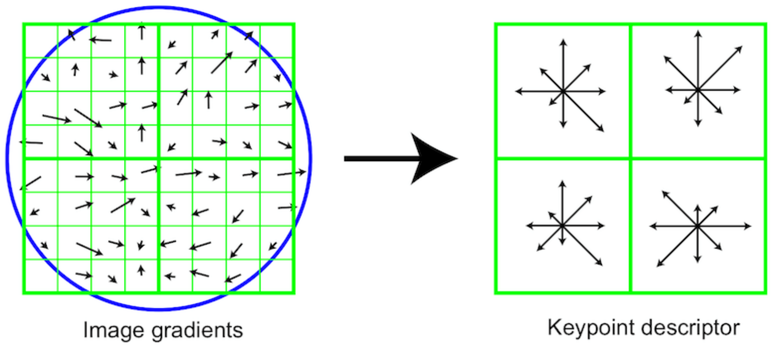
- Mountain
- Snow-capped mountain
- Cloud
- Blue sky
- Wild flowers
- Landscape
- Bushes
- Etc.

The vagueness of annotation is caused by human's perception subjectivity,
Besides, different people may catch different semantic meaning of the image

Past, Today and Tomorrow

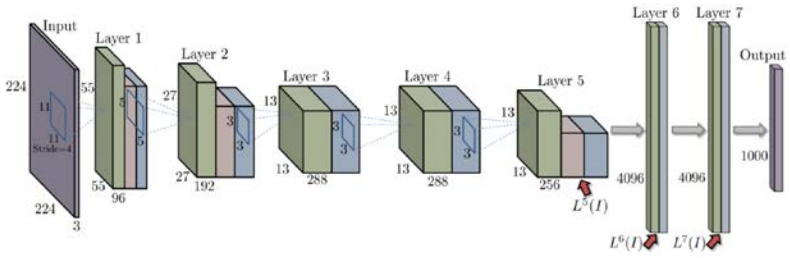
Past:

Classic methods

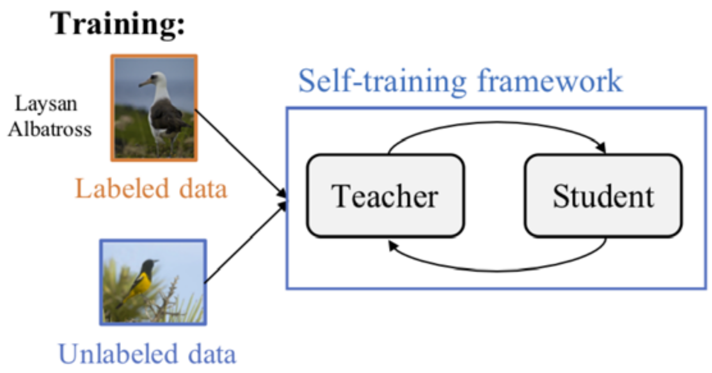


Today:

Deep-learning based



Future:
???



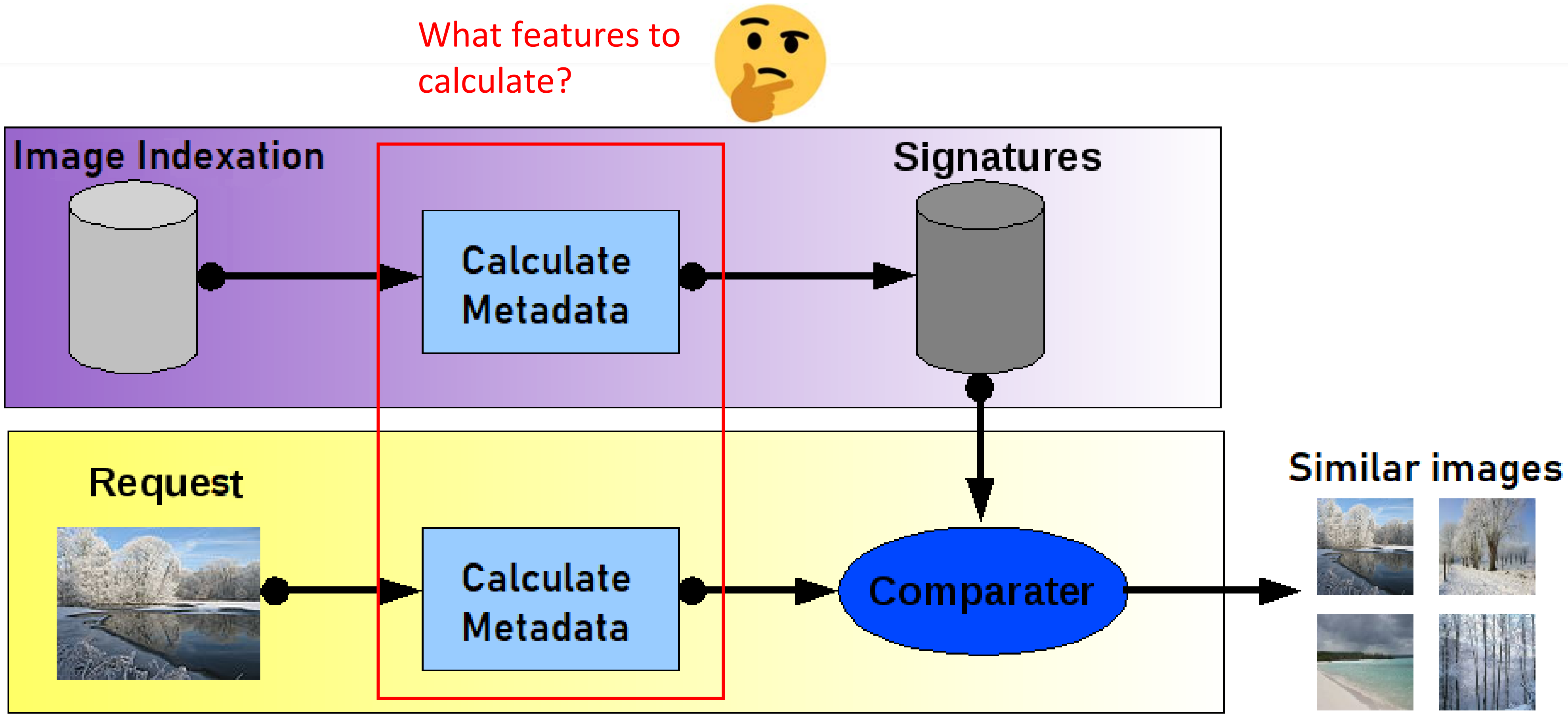
Yesterday (1990-2012): Unsupervised CBIR

The 1^{st} Decade (1990-2000)

1990-2000: Exploration Stage

- Color Histogram
- Texture Features
- Shape Features

CBIR System


What features to
calculate?
Image Indexation
Signatures
Calculate
Metadata
Request
Calculate
Metadata
Comparater
Similar images

Classical Methods (1990-2000)

An image distance measure compares the similarity of two images in various feature spaces such as color, texture, shape, and others

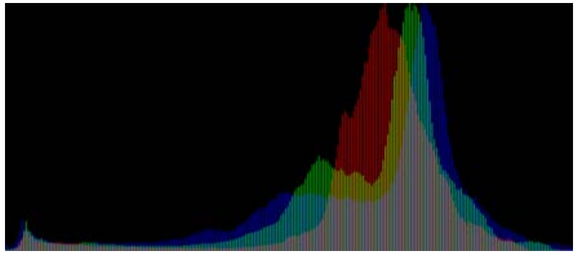
- Color: color histogram - the proportion of pixels holding specific values
- Texture: measures look for visual patterns and how they are spatially defined
- Shape: shape filters to identify given shapes of an image

Rui, Yong, Thomas S. Huang, and Shih-Fu Chang. "Image retrieval: Current techniques, promising directions, and open issues." *Journal of visual communication and image representation* 10, no. 1 (1999): 39-62.

Color Descriptor: Color Histogram

- A color histogram is a representation of the distribution of colors in an image
- A color histogram represents the number of pixels that have colors in each of a fixed list of color ranges, that span the image's color space.

A picture of a cat



Color histogram of the above cat picture with x-axis being RGB and y-axis being the frequency.

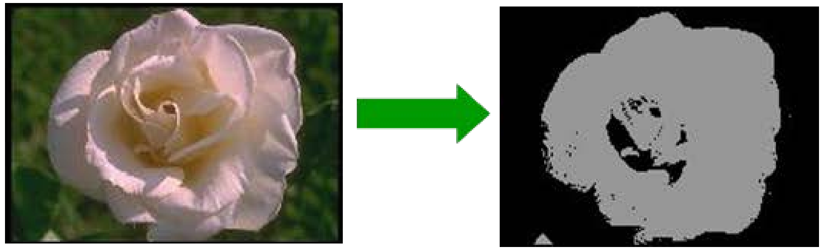
Color histogram from Wikipedia: https://en.wikipedia.org/wiki/Color_histogram

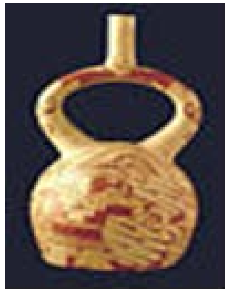
Texture Descriptor: Wavelet Transform

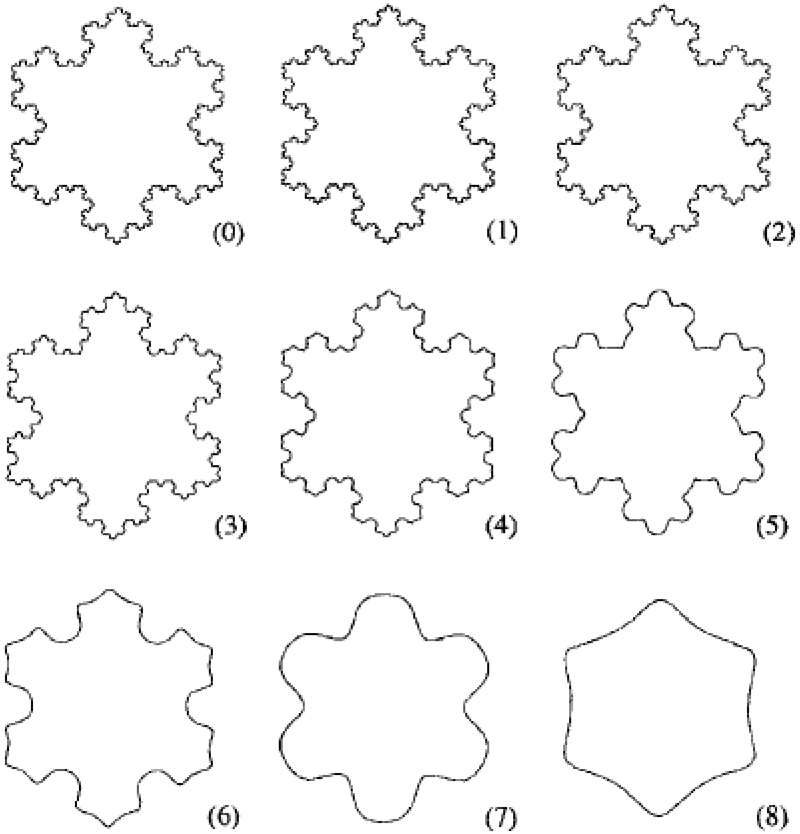
- A wavelet series is a representation of a square-integrable (real- or complex-valued) function by a certain orthonormal series generated by a wavelet
- Gabor wavelets, tree-structured wavelets

Wikipedia: https://en.wikipedia.org/wiki/Wavelet_transform

Shape Descriptor: Fourier/Wavelet Transform







- Fourier curve descriptor
- Wavelet curve descriptor

Wikipedia: https://en.wikipedia.org/wiki/Fourier_transform

The 2nd Decade (2000-2010)

2000-2010: More robust and invariant feature representations

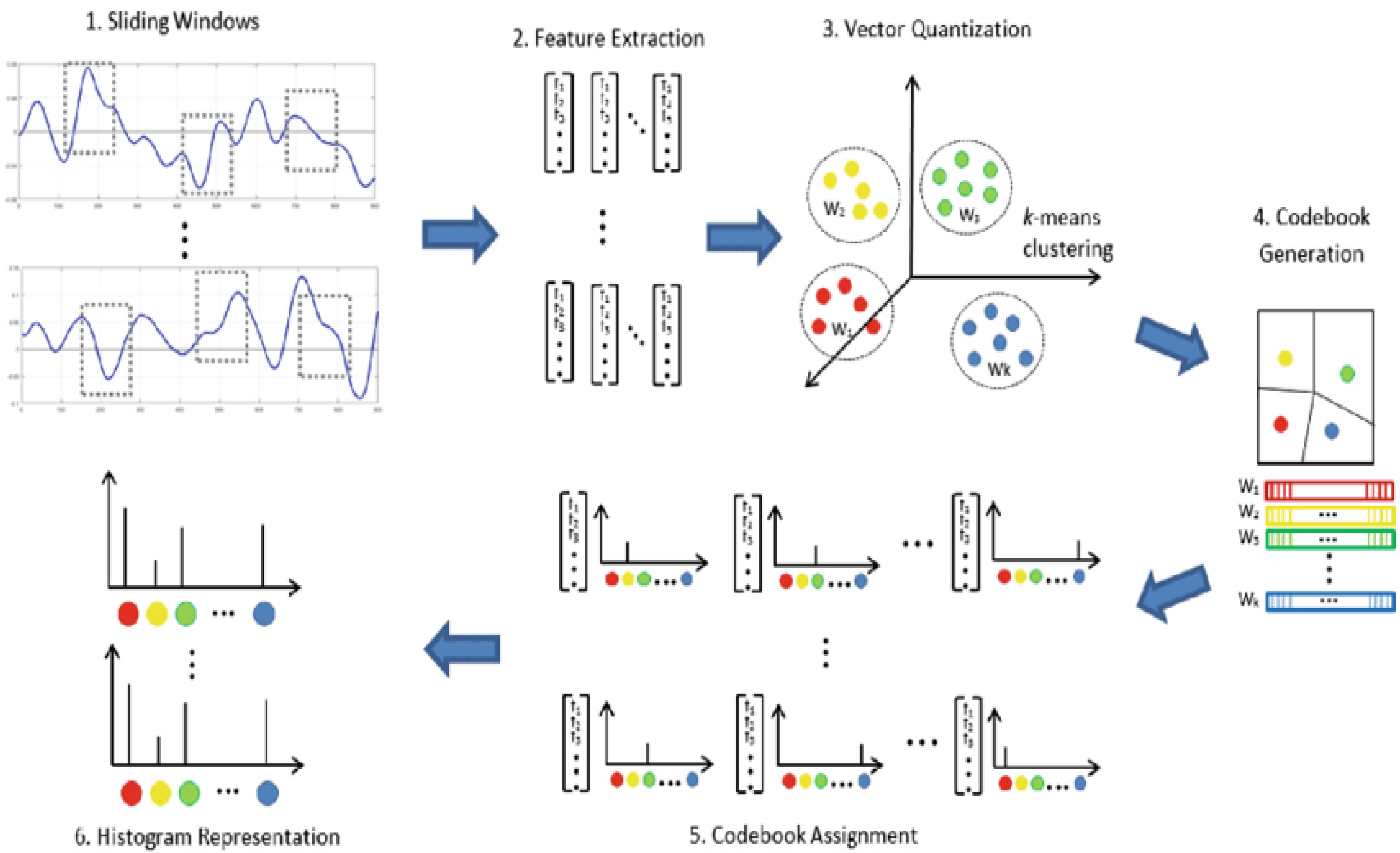
- BoW: An orderless document representation — only the counts of words matter
- SIFT: Invariant to uniform scaling, orientation and illumination changes
- HOG: Counts occurrences of gradient orientation in localized portions of an image

Improved Feature Descriptors

- **Robustness and invariance**
  - Bag-of-words (BoW) model
  - Scale-invariant feature transform (SIFT)
  - Histogram of Oriented Gradients (HoG)

Bag-of-Words model (BoW)

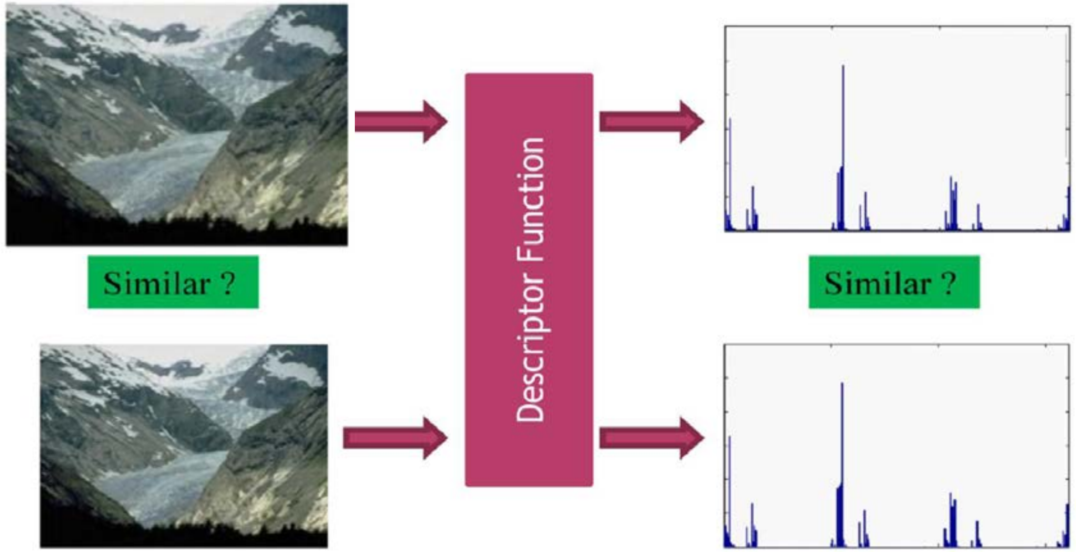
Algorithm:
- Sliding Windows
- Feature Extraction
- Vector Quantization
- Codebook Generation
- Codebook Assignment
- Histogram Representation



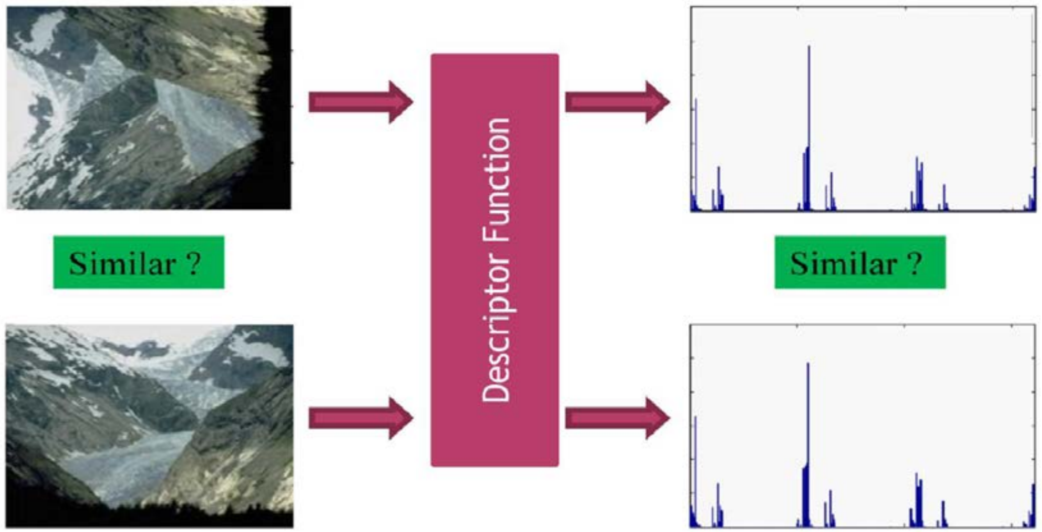
Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. "Understanding bag-of-words model: a statistical framework." *International Journal of Machine Learning and Cybernetics* 1.1-4 (2010): 43-52.

Desired Properties

- Invariance to scale



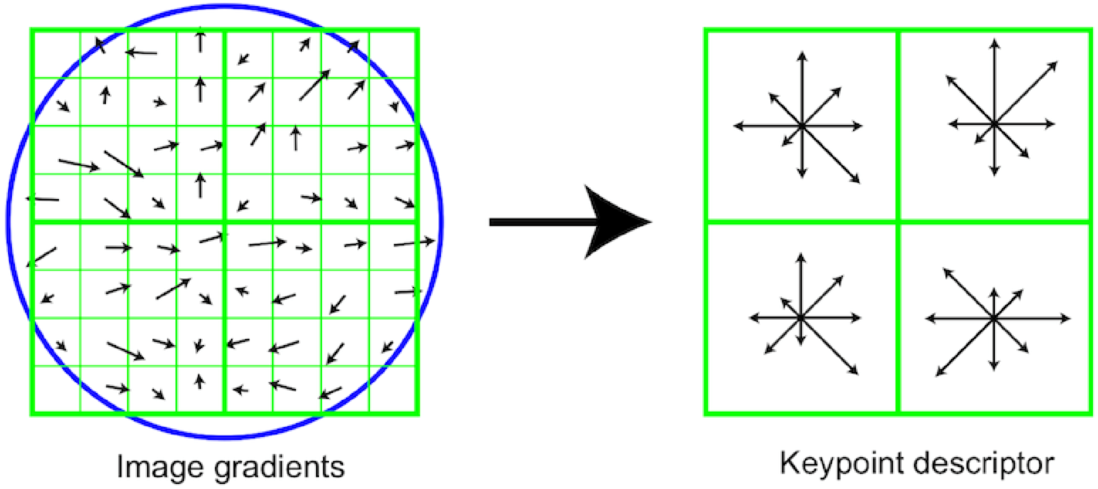
- Invariance to orientation



Scale-Invariant Feature Transform (SIFT)

SIFT Computation:

- Feature point (keypoint) detection
- Feature point localization
- Orientation assignment
- Feature descriptor generation

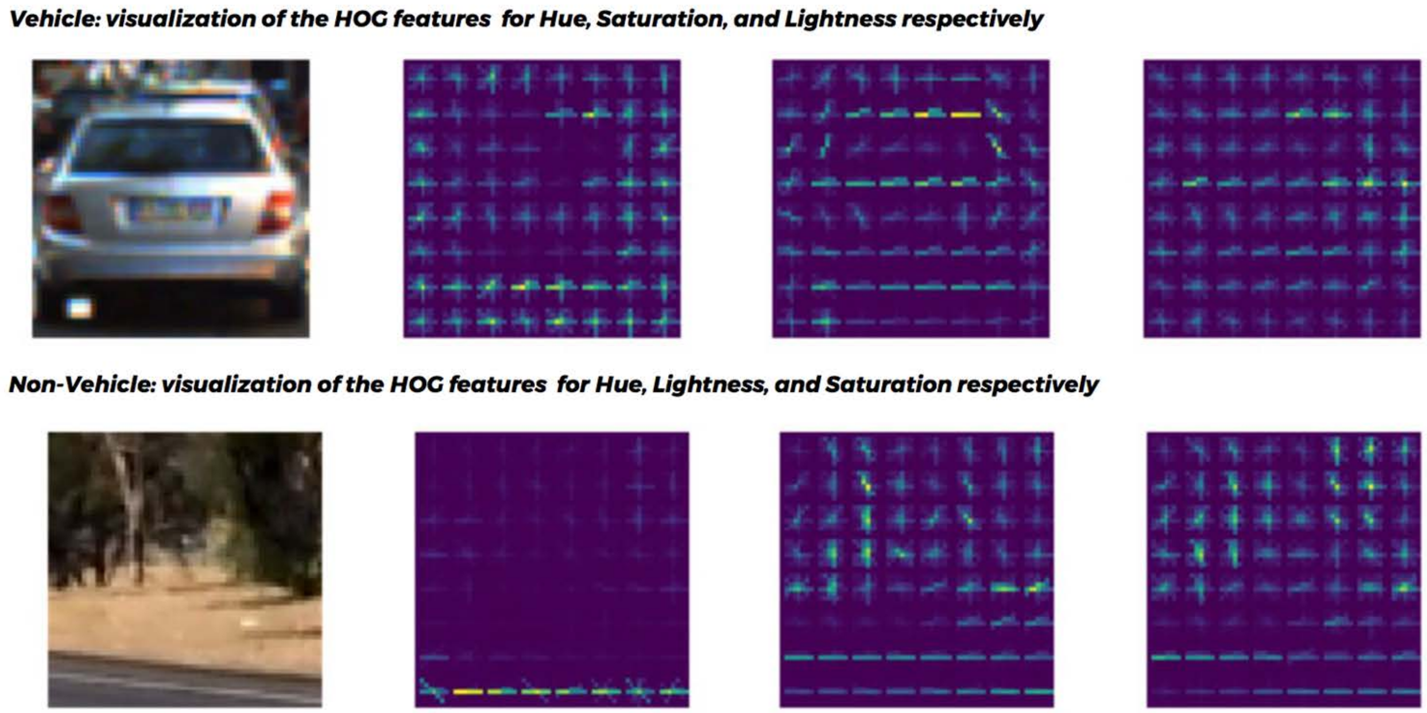


Lowe, G. "SIFT-the scale invariant feature transform." *Int. J* 2 (2004): 91-110.

Histogram of Oriented Gradients (HoG)

HoG Computation:

- Gradient computation
- Orientation binning
- Descriptor blocks
- Block normalization



Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *CVPR 2005*

Applications

- Starbucks Sign
- BMW Sign

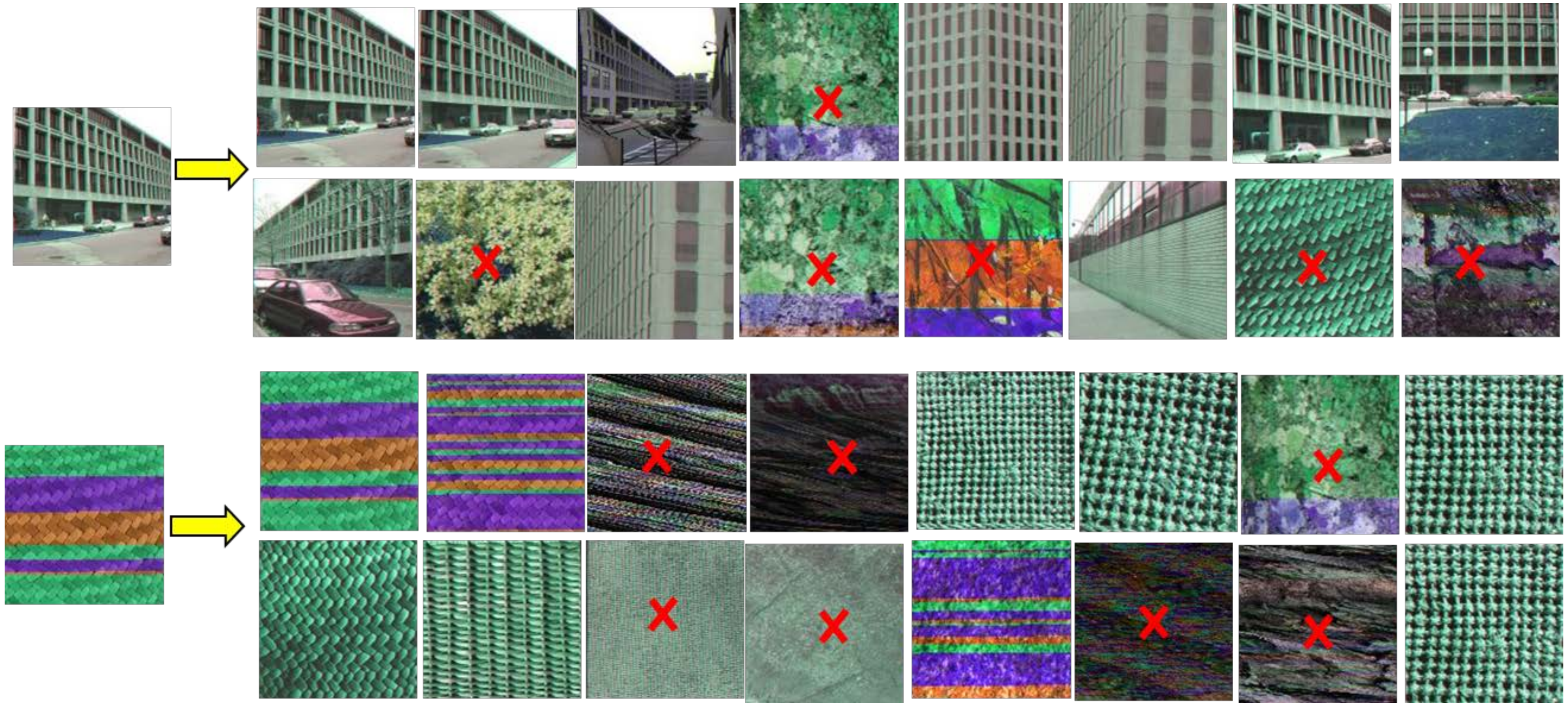
How to resolve the Gap between Image Pixels and Semantics?

Supervision!

Today (2012-Present): Supervised CBIR

A Form of “Supervision”

- Relevance Feedback (1998)



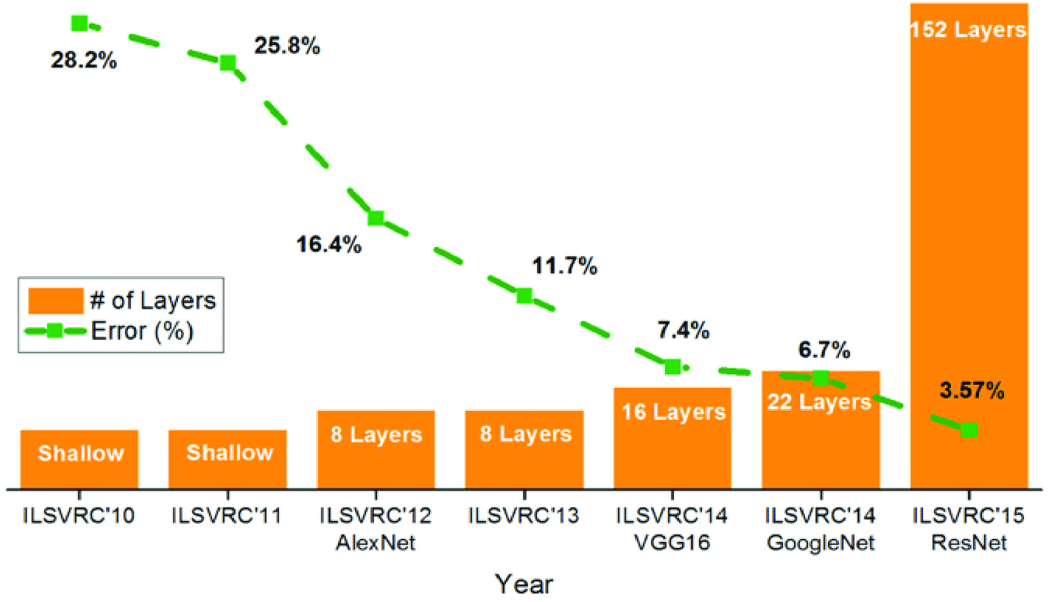
Rui, Yong, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. "Relevance feedback: A power tool for interactive content-based image retrieval." IEEE Transactions on circuits and systems for video technology 8, no. 5 (1998): 644-655.

Little Follow-up?

- **No mature machine learning methodology**
- **Lack of large-scale training datasets**
- **Lack of powerful features**
  - Domain knowledge (feature engineering)
  - Lack of robustness
  - Poor performance

Turning Point: 2012

ImageNet Dataset



ILSVRC Challenge

Datasets: CUB-200-2011 & Stanford-Cars

CUB-200-2011

- Caltech/UCSD
- 200 categories of birds
- Number of images: 11,788

Stanford-Cars

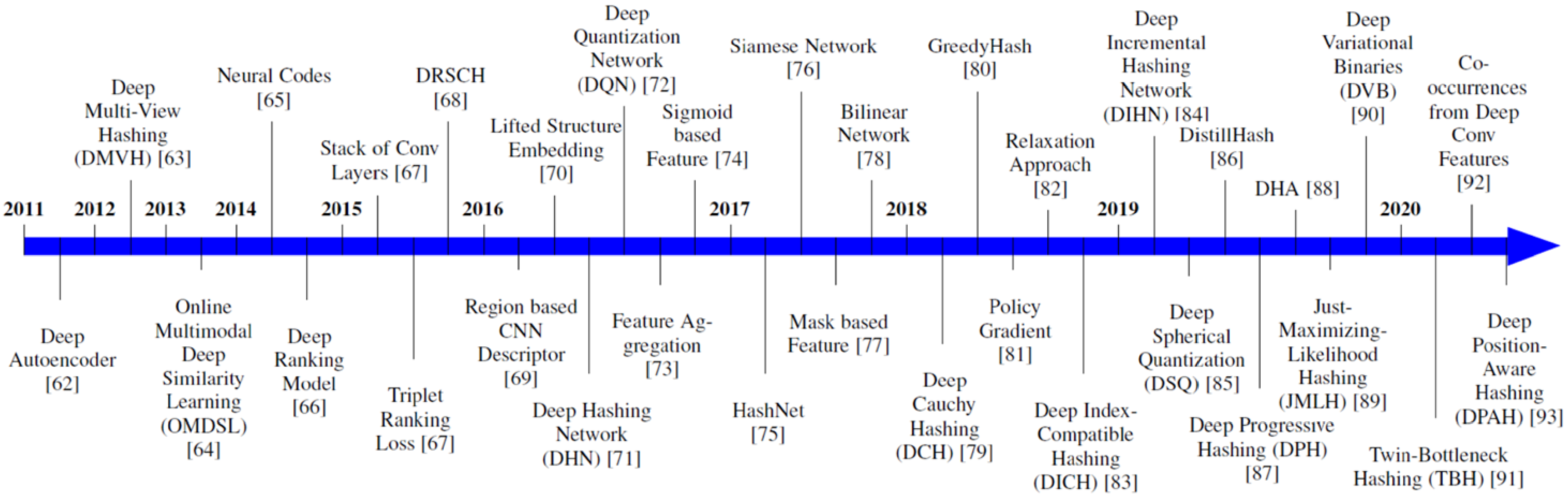
- 16,185 images made up of 196 classes
- Classes are typically at the level of *Make, Model, Year*, e.g., 2012 Tesla Model S

Two Major Changes in Last Decade

- **Large-scale labeled datasets**
  - CUB-200-2011 (2011) – fine-grained bird retrieval dataset
  - CAR-196 (2013) – fine-grained car retrieval dataset
  - Market-1501 (2015) – person re-identification dataset
  - Stanford Online Shopping (2016) – a variety of online shopping instances
- **Resurgence of neural networks**
  - Pro: generalizability (can handle a large amount of data) and superior performance
  - Cons: black box, adversarial attacks, high computational cost

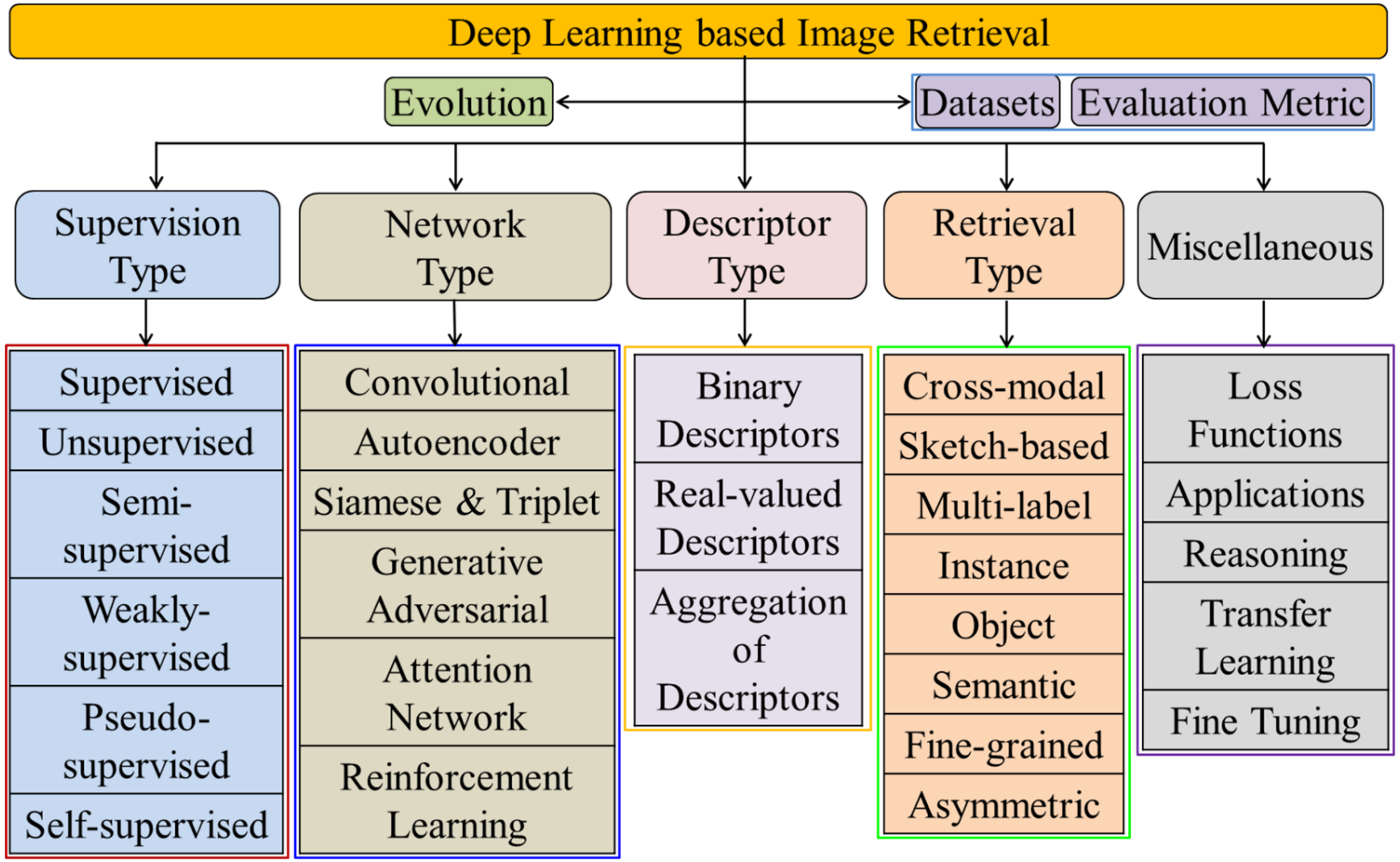
Evolution of DL-based Retrieval Methods

Chronological Overview



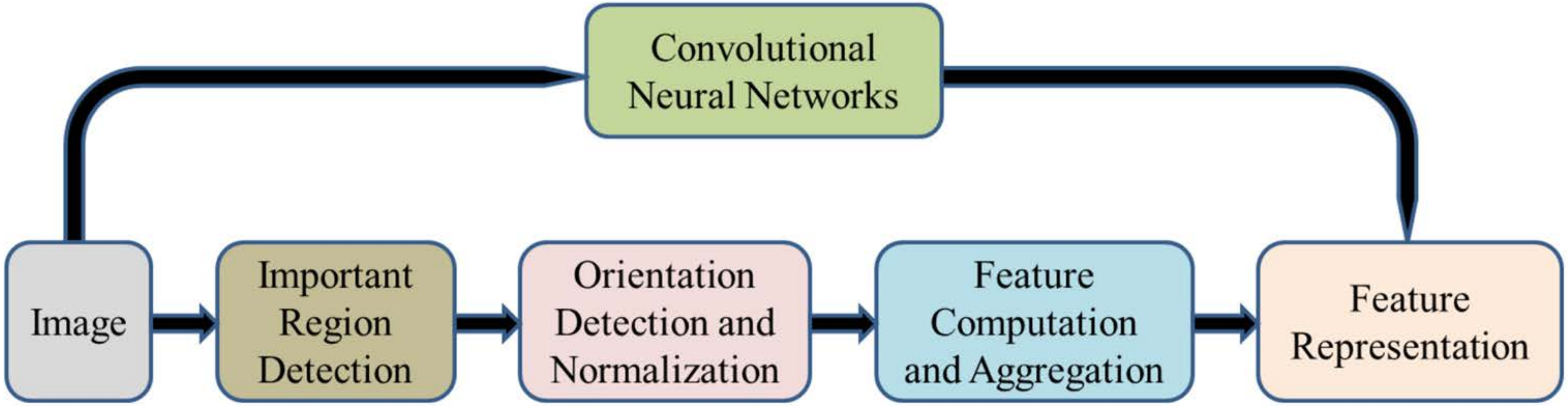
Dubey, Shiv Ram. "A Decade Survey of Content Based Image Retrieval using Deep Learning." *arXiv preprint arXiv:2012.00641* (2020).

Categorization of DL-based Image Retrieval Research



Dubey, Shiv Ram. "A Decade Survey of Content Based Image Retrieval using Deep Learning." *arXiv preprint arXiv:2012.00641* (2020).

Deep vs. Traditional Features



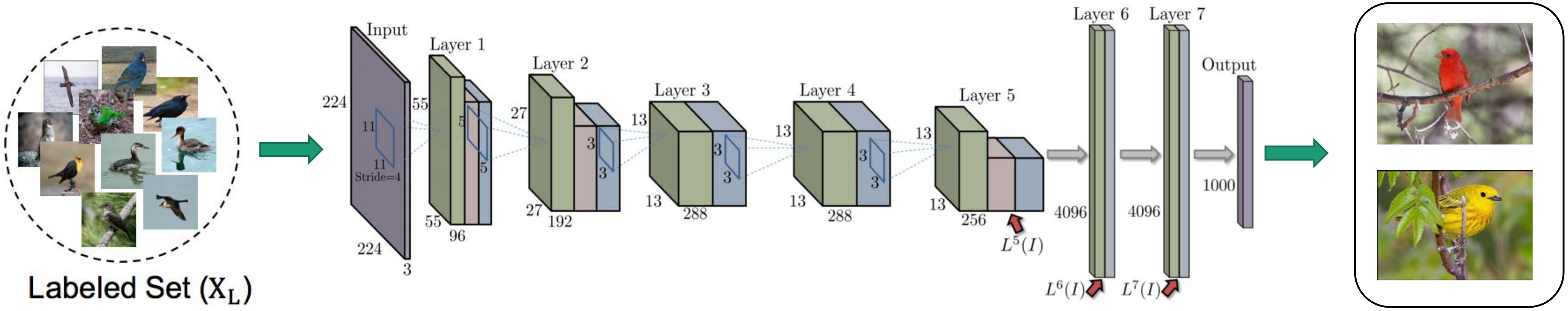
The pipeline of state-of-the-art feature representation is replaced by the CNN based feature representation with increased discriminative ability and robustness

Dubey, Shiv Ram. "A Decade Survey of Content Based Image Retrieval using Deep Learning." *arXiv preprint arXiv:2012.00641* (2020).

Learning Features and Distances from Data

Key: Design of the Loss Function

Similar or not?

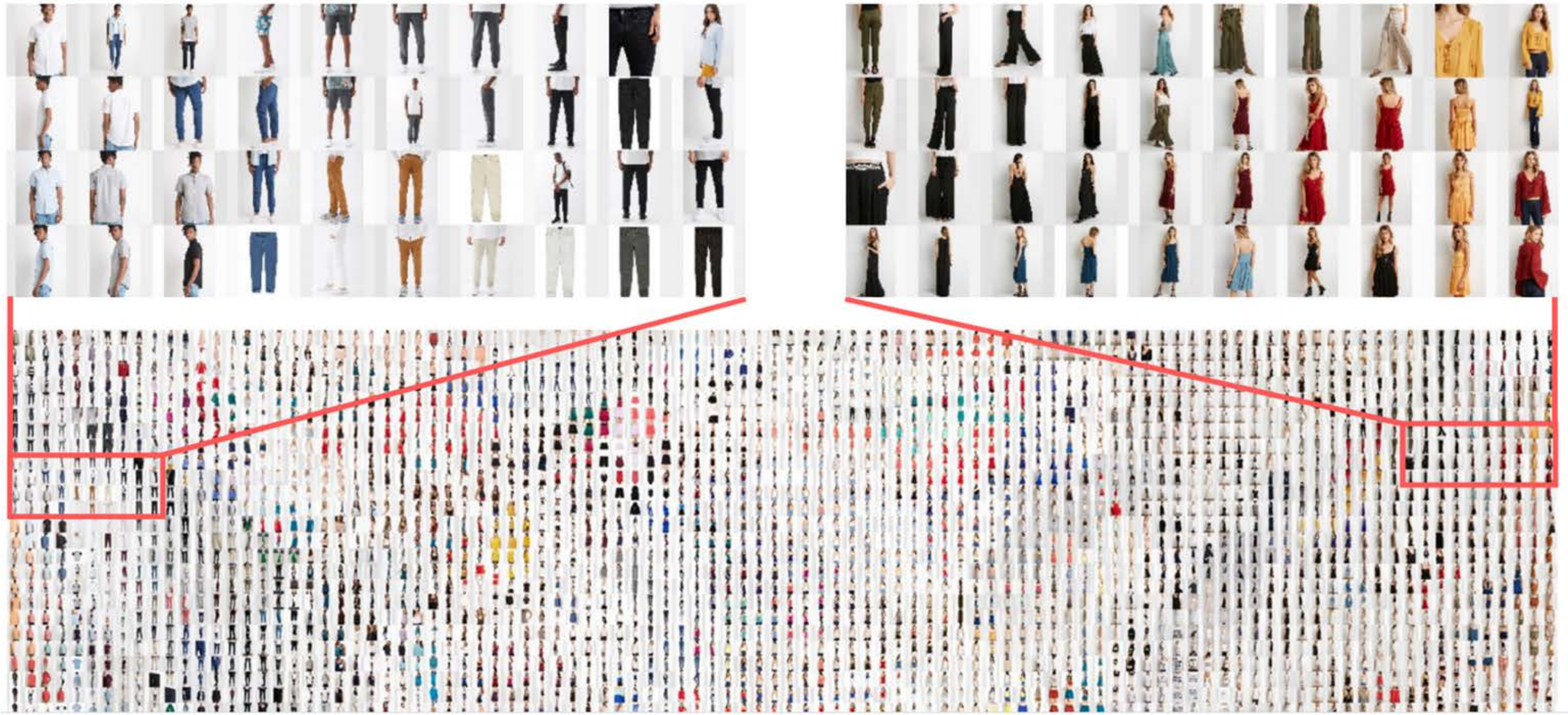


Input: Labeled, unlabeled, partially labeled

Various architectural design choices: Siamese, U-Net, Skip-connections etc.,

Loss/Metric design choices

Learning Regularized Embedding Space



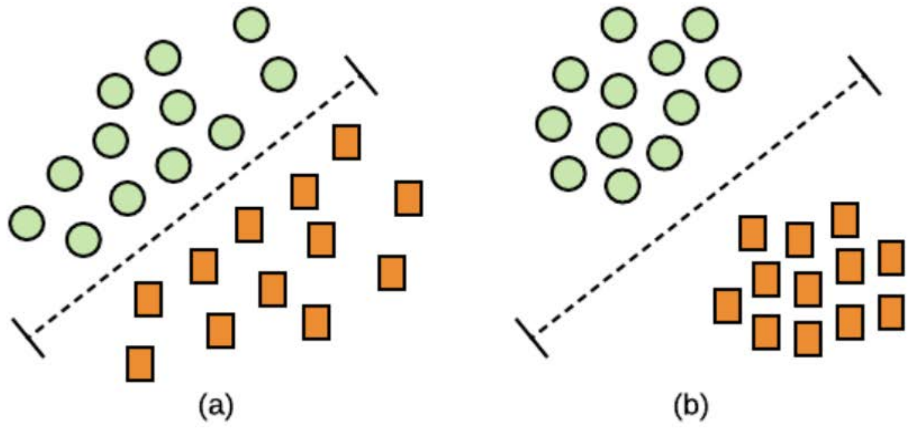
Design of Loss Functions

- **Pair & Proxy-based Loss**
  - Pair-based Loss: Contrastive & Triplet Loss
  - Pair-based vs Proxy-based
- **Ranking-based Loss**
  - Learning to rank: Fast-AP & Smooth-AP
  - Pair-based vs Ranking-based

Contrastive Loss & Triplet Loss

Take two input samples: similar or dissimilar

Goal of contrast loss: push similar samples closer and push dissimilar samples further away

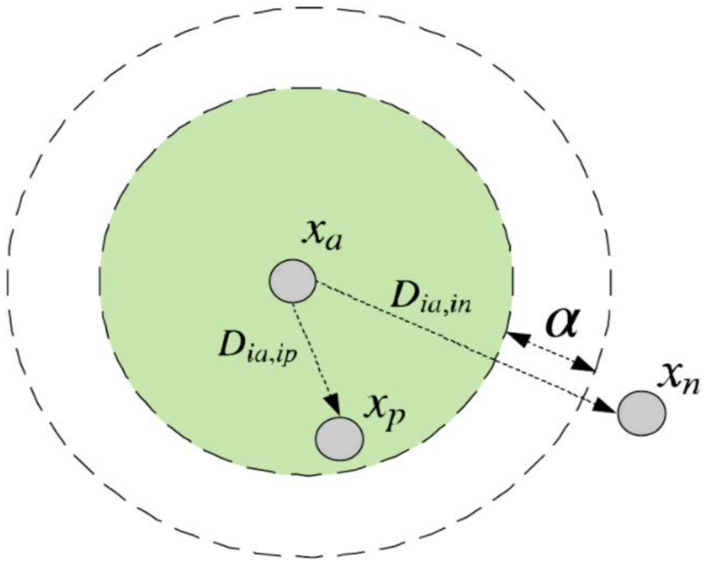


(a) Separable Features (b) Discriminative Features

$$L_{contrastive} = [d_p - m_{pos}]_+ + [m_{neg} - d_n]_+$$

Take three input samples: anchor, positive and negative ones

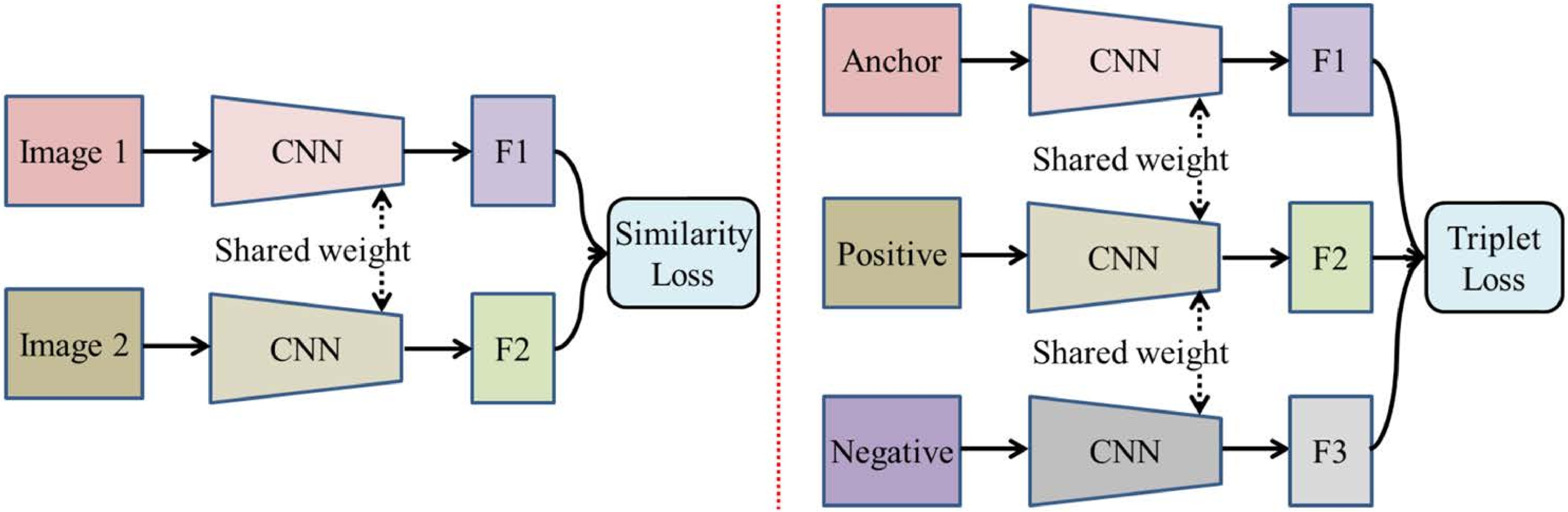
Goal of triple loss: push the anchor closer to the positive one and far away from the negative one



$$L_{triplet} = [d_{ap} - d_{an} + m]_+$$

Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

Contrastive Loss & Triplet Loss



(a) Siamese Network

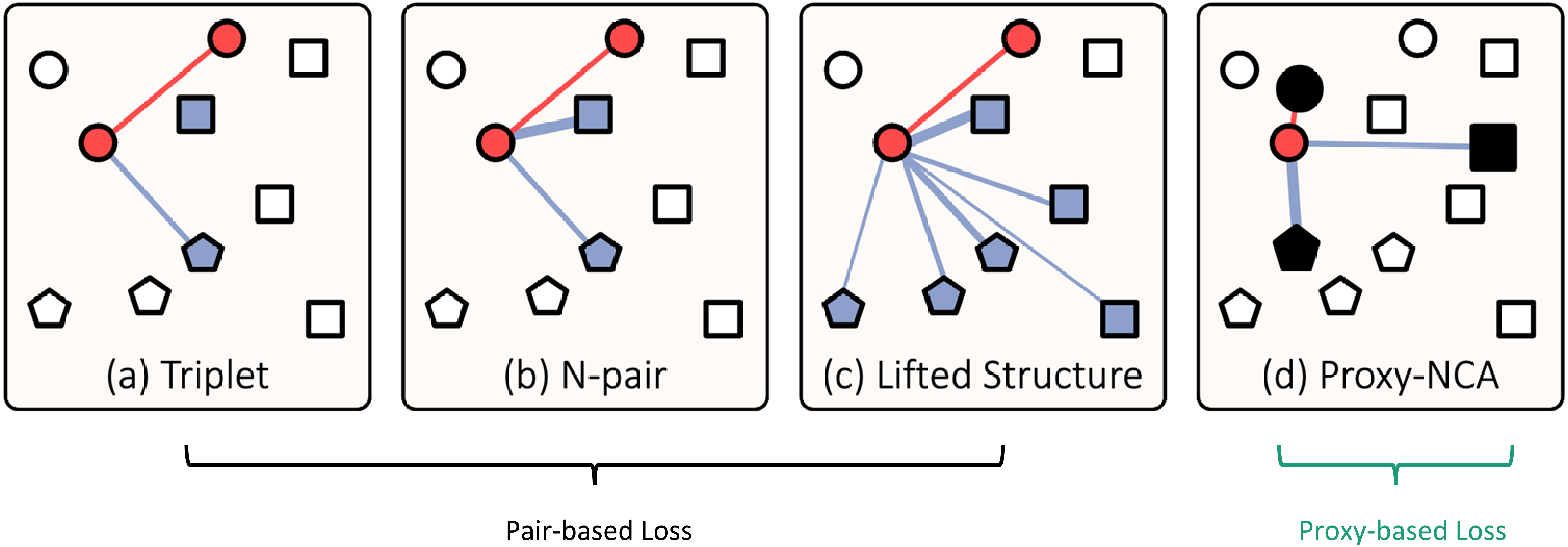
(b) Triplet Network

Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

Pair-based vs Proxy-based Loss

Pair-based Loss: Sampling matters!

Low sampling complexity!



1. Sohn K. Improved deep metric learning with multi-class n-pair loss objective[J]. Advances in neural information processing systems, 2016, 29: 1857-1865.
2. Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and SilvioSavarese. Deep metric learning via lifted structured feature embedding. CVPR 2016
3. Yair Movshovitz-Attias, Alexander Toshev, Thomas K Le-ung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. ICCV 2017

Pair-based vs Proxy-based Loss

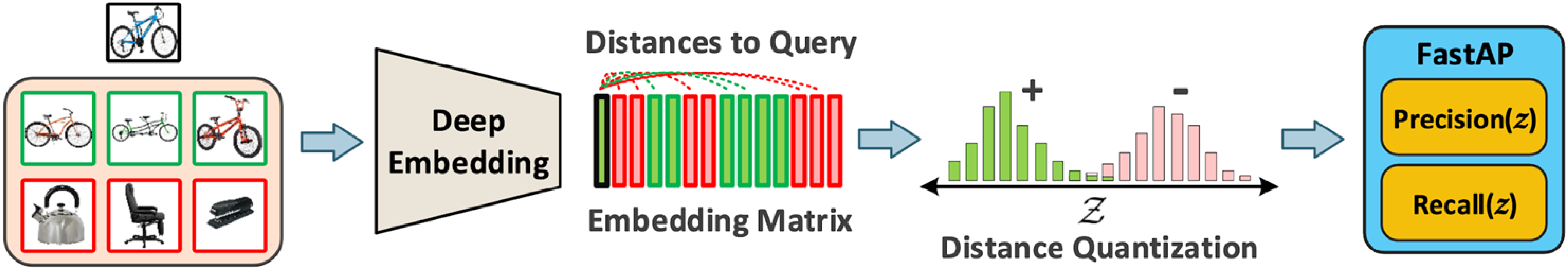
Type	Loss	Training Complexity
Proxy	Proxy-Anchor (Ours)	$O(MC)$
	Proxy-NCA [21]	$O(MC)$
	SoftTriple [23]	$O(MCU^2)$
Pair	Contrastive [2, 4, 9]	$O(M^2)$
	Triplet (Semi-Hard) [25]	$O(M^3/B^2)$
	Triplet (Smart) [10]	$O(M^2)$
	N-pair [27]	$O(M^3)$
	Lifted Structure [29]	$O(M^3)$

Training complexity comparison

Kim S, Kim D, Cho M, et al. Proxy Anchor Loss for Deep Metric Learning. CVPR 2020

Ranking-based Loss (1)

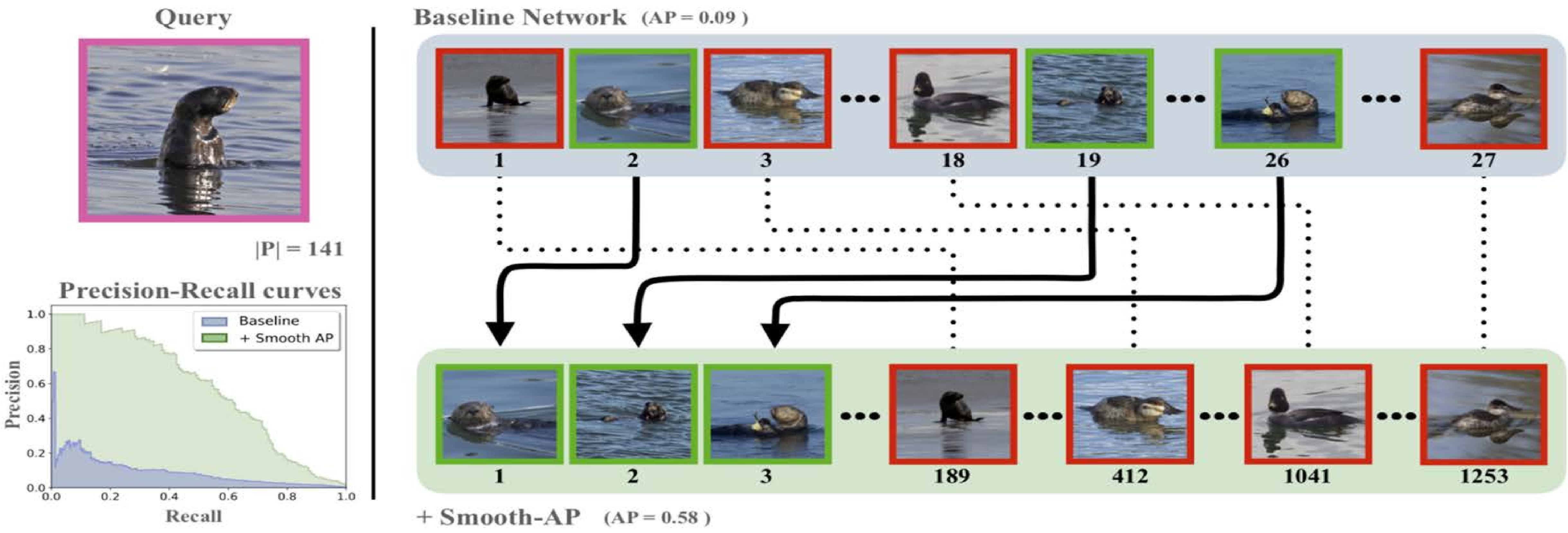
Example: Fast AP
- A deep metric learning method to rank
- Optimize both Precision and Recall for better Average Precision (AP)



Cakir F, He K, Xia X, et al. Deep metric learning to rank. CVPR 2019

Ranking-based Loss (2)

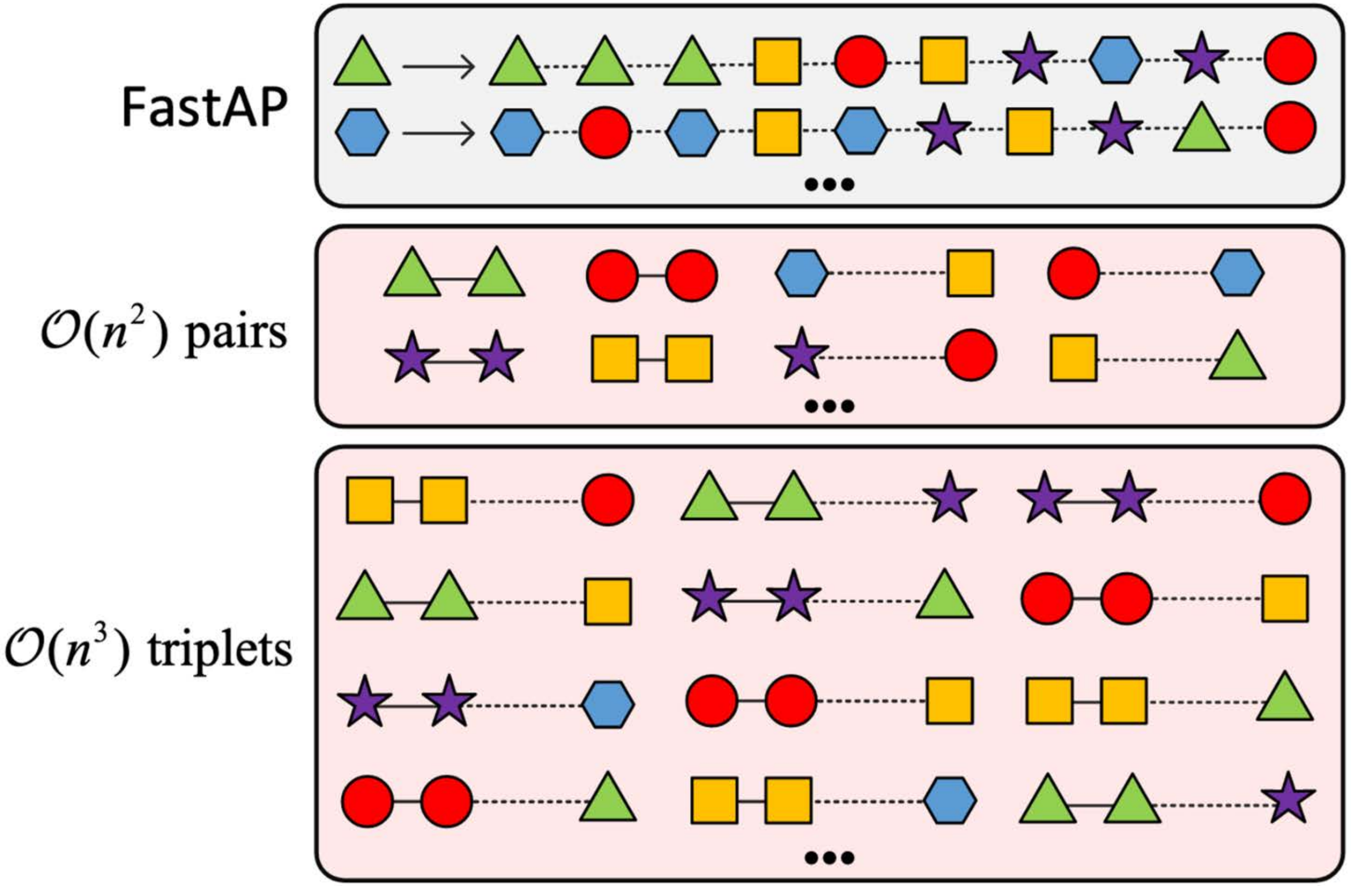
Example: Smooth-AP
- Smoothing the path towards large-scale image retrieval



Ranked retrieval sets before (top) and after (bottom) applying Smooth-AP.

Brown A, Xie W, Kalogeiton V, et al. Smooth-AP: Smoothing the path towards large-scale image retrieval. ECCV 2020

Comparison: Pair-based vs Ranking-based Loss



Explicitly optimize ranking

Pros:

- Directly optimize rank
- Use more samples

Cons:

- Non-differentiable
- Require more memory

Cakir F, He K, Xia X, et al. Deep metric learning to rank. CVPR 2019

Performance Comparison

		Concatenated (512-dim)			Separated (128-dim)		
		P@1	RP	MAP@R	P@1	RP	MAP@R
	Pretrained	46.89	13.77	5.91	43.27	13.37	5.64
Pair-based	Contrastive	81.78 ± 0.43	35.11 ± 0.45	24.89 ± 0.50	69.80 ± 0.38	27.78 ± 0.34	17.24 ± 0.35
	Triplet	79.13 ± 0.42	33.71 ± 0.45	23.02 ± 0.51	65.68 ± 0.58	26.67 ± 0.36	15.82 ± 0.36
	NT-Xent	80.99 ± 0.54	34.96 ± 0.38	24.40 ± 0.41	68.16 ± 0.36	27.66 ± 0.23	16.78 ± 0.24
Proxy-based	ProxyNCA	83.56 ± 0.27	35.62 ± 0.28	25.38 ± 0.31	73.46 ± 0.23	28.90 ± 0.22	18.29 ± 0.22
Pair-based	Margin	81.16 ± 0.50	34.82 ± 0.31	24.21 ± 0.34	68.24 ± 0.35	27.25 ± 0.19	16.40 ± 0.20
Classification	Margin / class	80.04 ± 0.61	33.78 ± 0.51	23.11 ± 0.55	67.54 ± 0.60	26.68 ± 0.40	15.88 ± 0.39
	N. Softmax	83.16 ± 0.25	36.20 ± 0.26	26.00 ± 0.30	72.55 ± 0.18	29.35 ± 0.20	18.73 ± 0.20
Pair-based	CosFace	$\mathbf{85.52 \pm 0.24}$	37.32 ± 0.28	27.57 ± 0.30	$\mathbf{74.67 \pm 0.20}$	29.01 ± 0.11	18.80 ± 0.12
	ArcFace	85.44 ± 0.28	37.02 ± 0.29	27.22 ± 0.30	72.10 ± 0.37	27.29 ± 0.17	17.11 ± 0.18
Ranking-based	FastAP	78.45 ± 0.52	33.61 ± 0.54	23.14 ± 0.56	65.08 ± 0.36	26.59 ± 0.36	15.94 ± 0.34
Pair-based	SNR	82.02 ± 0.48	35.22 ± 0.43	25.03 ± 0.48	69.69 ± 0.46	27.55 ± 0.25	17.13 ± 0.26
	MS	85.14 ± 0.29	$\mathbf{38.09 \pm 0.19}$	$\mathbf{28.07 \pm 0.22}$	73.77 ± 0.19	$\mathbf{29.92 \pm 0.16}$	$\mathbf{19.32 \pm 0.18}$
Mining	MS+Miner	83.67 ± 0.34	37.08 ± 0.31	27.01 ± 0.35	71.80 ± 0.22	29.44 ± 0.21	18.86 ± 0.20
Proxy-based	SoftTriple	84.49 ± 0.26	37.03 ± 0.21	27.08 ± 0.21	73.69 ± 0.21	29.29 ± 0.16	18.89 ± 0.16

Performance comparison on Cars-196 dataset

Musgrave K, Belongie S, Lim S N. A metric learning reality check. ECCV 2020

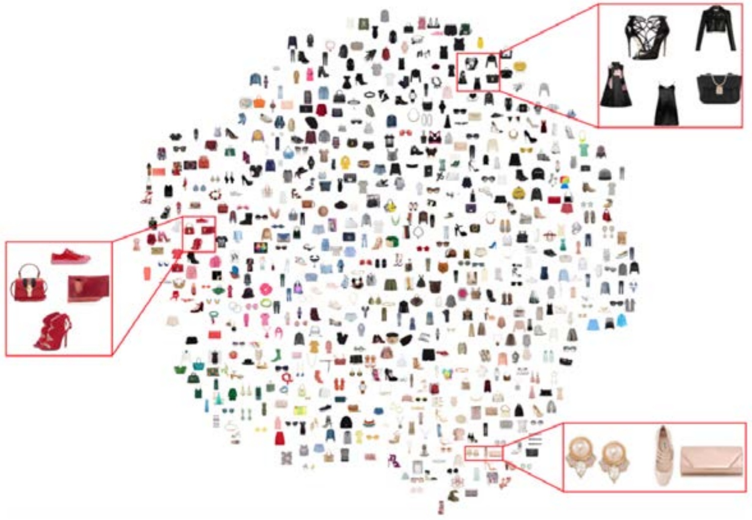
Tomorrow (Next Decade):
Back to Weak Supervision and Practical Applications

Two Main Directions

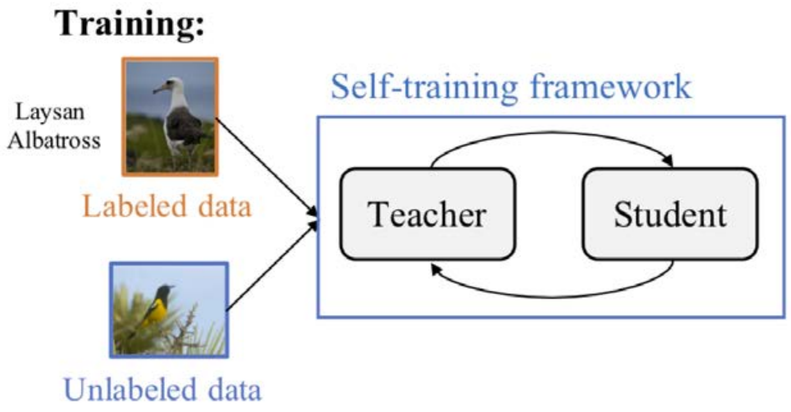
- **Push the envelop of deep learning (DL)**
- **Real world applications**
- **Towards green machine learning**
- **Cross-domain knowledge structure**
  - **Weakly-supervised learning**

Push the Envelop of DL

- **Relational Reasoning**
  - Model relationships between samples
  - Reason users' preferences



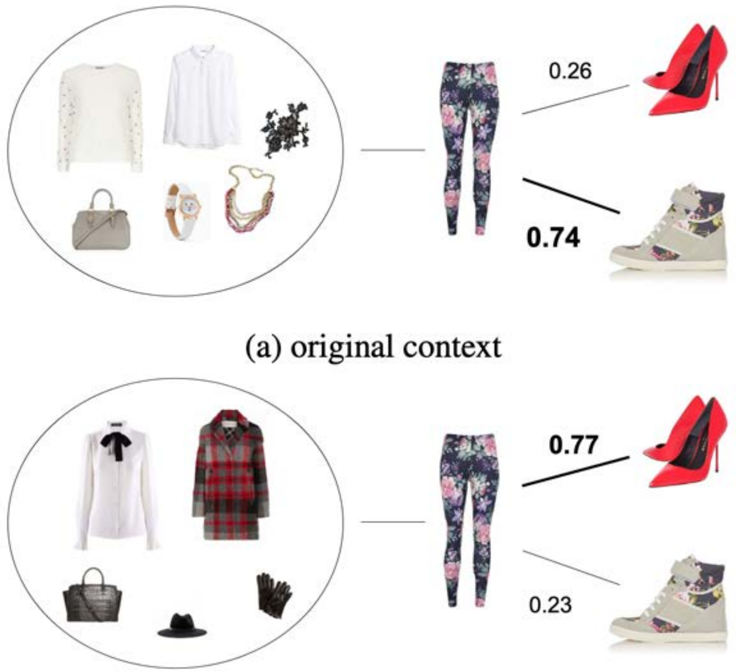
- **Leverage Unlabeled Data**
  - Data annotation is expensive
  - Continual learning/life-long learning
  - Domain difference between datasets
  - Noisy samples and outliers



Relational Reasoning:
Fashion Compatibility Recommendation

- Goal 1: Recommend for a partial outfit

- Goal 2: Is the outfit valid?



(a) original context

(b) new context

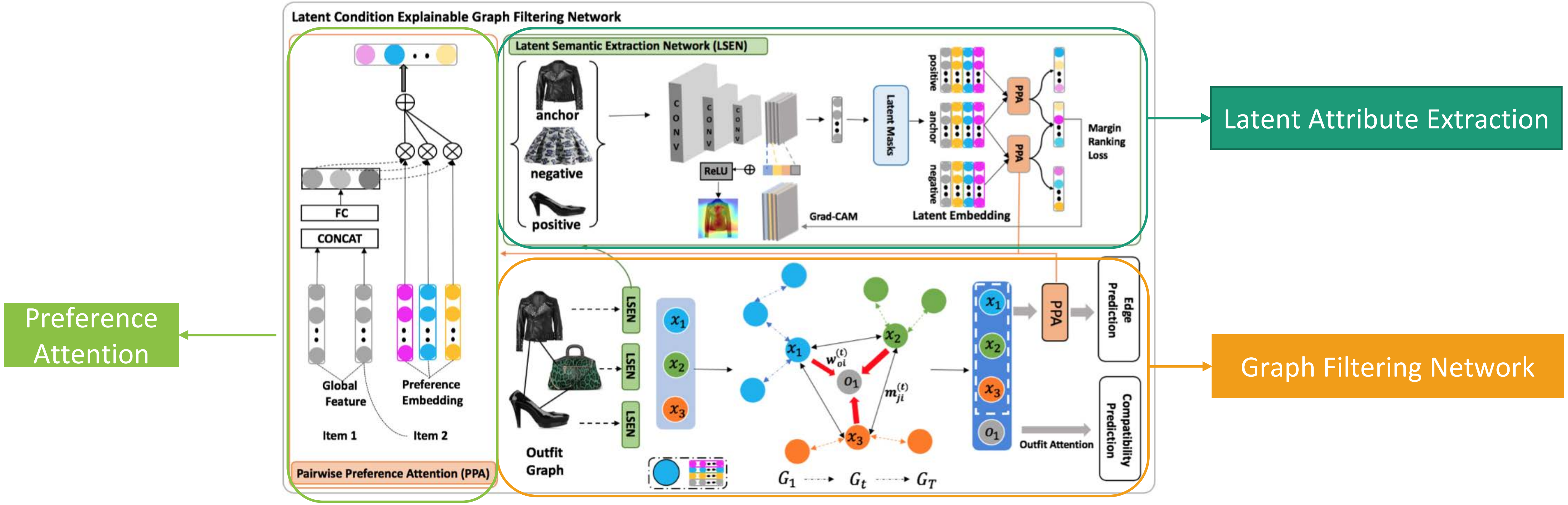
➢ Model the dependencies between items

➢ Context matters for compatibility prediction

J Duan, C.-C. Jay Kuo. Fashion Compatibility Recommendation via Unsupervised Metric Graph Learning. SCMLS 2020

Overview of Proposed Framework

- Attribute Aware Explainable Graph Network (AAEG)



J Duan, C.-C. Jay Kuo. Fashion Compatibility Recommendation via Unsupervised Metric Graph Learning. SCMLS 2020

Results: Quantitative Evaluation

Method	FITB ACC	Compat. AUC
Siamese Net [135]	54.4%	0.85
Bi-LSTM [45]	64.9%	0.94
TA-CSN [132]	65.0%	0.93
SCE-Net [125]	60.8%	0.90
CA-GCN (wo /ctx) [30]	41.7%	0.71
CA-GCN (w /ctx) [30]	83.1%	0.99
Ours (wo/ ctx)	62.1%	0.93
Ours (w/ ctx)	87.3%	0.99
Ours + Outfit(w/ ctx)	**89.3%**	**0.99**

Fill-in-the-Blank (FITB)

Compatibility of the Outfit (Compat.)

J Duan, C.-C. Jay Kuo. Fashion Compatibility Recommendation via Unsupervised Metric Graph Learning. SCMLS 2020

Leverage Unlabeled Data

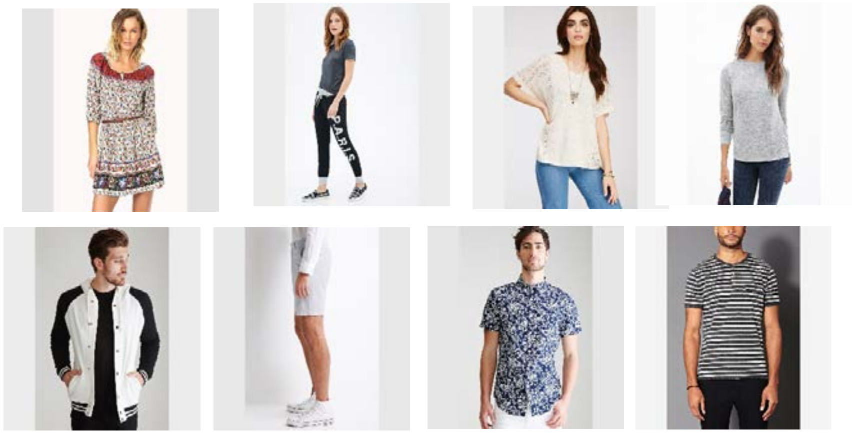
- Existing methods require pairwise annotation

Same Class

Different Class

…

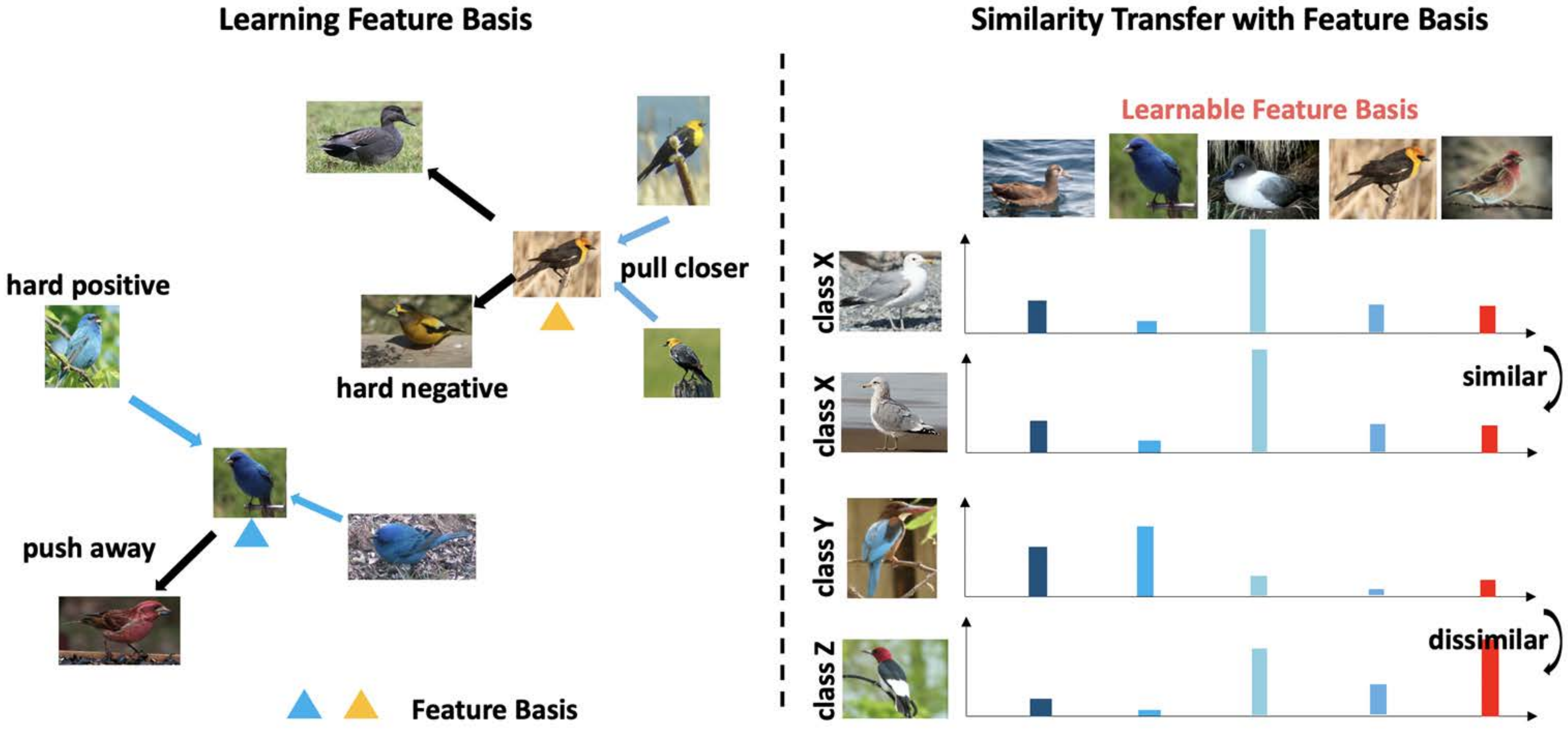
- Un-annotated data has not been leveraged



…

Goal: Leverage un-annotated data to improve deep metric learning

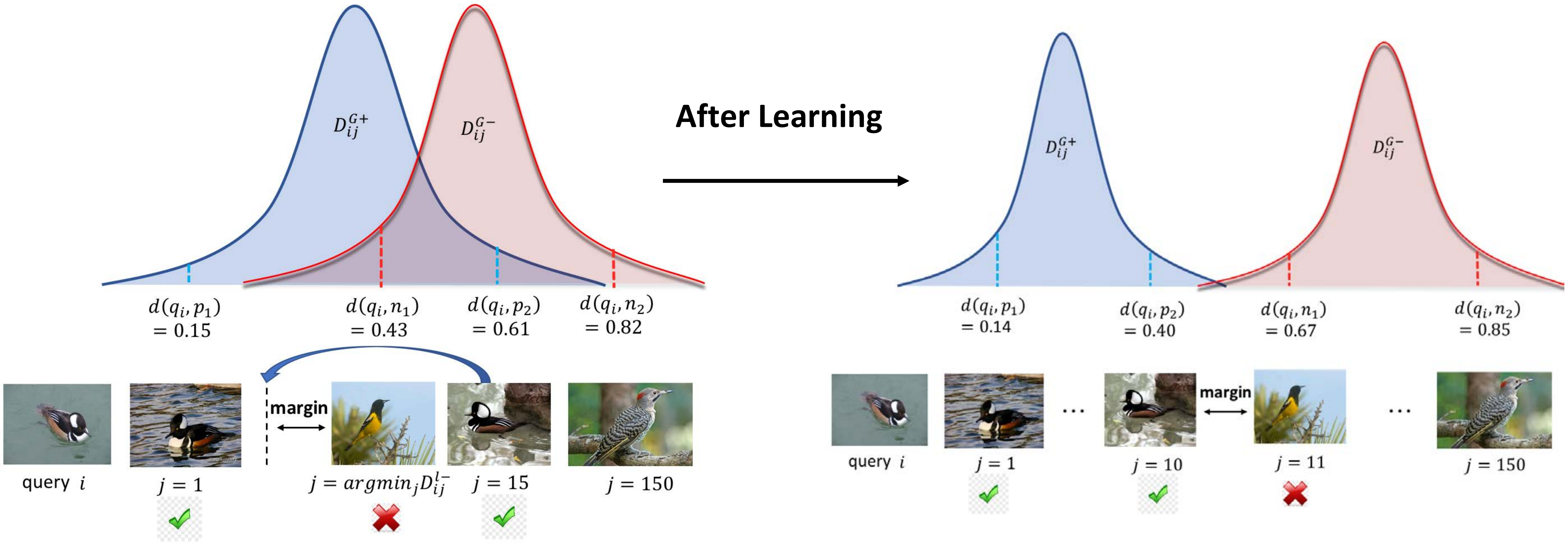
Feature Basis Learning



Similarity Distribution Loss

Goal: reducing overlap between distributions

1. Maximize distance between two means
2. Reduce variances of two distributions



Performance Comparison

Methods	Frwk	Init	Arc / Dim	CUB-200-2011			Cars-196		
				MAP@R	RP	P@1	MAP@R	RP	P@1
Contrastive [10]	[19]	ImageNet	BN / 512	26.53	37.24	68.13	24.89	35.11	81.78
Triplet [29]	[19]	ImageNet	BN / 512	23.69	34.55	64.24	23.02	33.71	79.13
ProxyNCA [18]	[19]	ImageNet	BN / 512	24.21	35.14	65.69	25.38	35.62	83.56
N. Softmax [35]	[19]	ImageNet	BN / 512	25.25	35.99	65.65	26.00	36.20	83.16
CosFace [25, 26]	[19]	ImageNet	BN / 512	26.70	37.49	67.32	27.57	37.32	85.52
FastAP [3]	[19]	ImageNet	BN / 512	23.53	34.20	63.17	23.14	33.61	78.45
MS+Miner [27]	[19]	ImageNet	BN / 512	26.52	37.37	67.73	27.01	37.08	83.67
Proxy-Anchor[1] [15]	[15]	ImageNet	R50 / 512	-	-	69.9	-	-	87.7
Proxy-Anchor[2] [15]	[19]	ImageNet	R50 / 512	25.56	36.38	66.04	30.70	40.52	86.84
ProxyNCA++ [22]	[22]	ImageNet	R50 / 2048	-	-	72.2	-	-	90.1
Mutual-Info [1]	[1]	ImageNet	R50 / 2048	-	-	69.2	-	-	89.3
Contrastive [10] (T_1)	[19]	ImageNet	R50 / 512	25.02	35.83	65.28	25.97	36.40	81.22
Contrastive [10] (T_2)	[19]	SwAV	R50 / 512	29.29	39.81	71.15	31.73	41.15	88.07
SLADE (Ours) (S_1)	[19]	ImageNet	R50 / 512	29.38	40.16	68.92	31.38	40.96	85.8
SLADE (Ours) (S_2)	[19]	SwAV	R50 / 512	**33.59**	**44.01**	**73.19**	**36.24**	**44.82**	**91.06**
MS [27] (T_3)	[19]	ImageNet	R50 / 512	26.38	37.51	66.31	28.33	38.29	85.16
MS [27] (T_4)	[19]	SwAV	R50 / 512	29.22	40.15	70.81	33.42	42.66	89.33
SLADE (Ours) (S_3)	[19]	ImageNet	R50 / 512	30.90	41.85	69.58	32.05	41.50	87.38
SLADE (Ours) (S_4)	[19]	SwAV	R50 / 512	**33.90**	**44.36**	**74.09**	**37.98**	**46.92**	**91.53**

J Duan, C.-C. Jay Kuo. SLADE: A Self-Training Framework For Distance Metric Learning. Arxiv Preprint 2020.

Real World Applications

- **Search through exemplary image/audio/video**
  - Identifying unknown plants, insects, animals, etc.
  - Preference search, e.g., songs, clothes, etc.
  - Surveillance, e.g., person re-identification, vehicle re-identification
- **Challenges**
  - User-satisfied performance (demanding a lot of data)
  - An engineering problem which is more suitable for industry

Concerns with Deep Learning

- **Not suitable for academic research**
  - Demanding heavy resources
    - Computing resource (GPU)
    - Data collection/labeling cost
  - Engineering fine-tuning
    - Blackbox tools – discouraging original thinking
- **Previous examples**
  - Computer graphics and SIGGRAPH
  - Image/video coding and standard meetings

An Alternative?

- **Green Machine Learning**
  - Decouple “feature extraction” and “decision” again
    - Feature extraction – unsupervised, statistics-based, signal processing (filter banks)
    - Decision – classification, regression, etc.
  - Unique characteristics
    - Low power consumption in both training and testing
    - Small model sizes
    - Suitable for edge/mobile devices
    - Also, beneficial to carbon footprint reduction in cloud servers

Example of Green Learning: DefakeHop

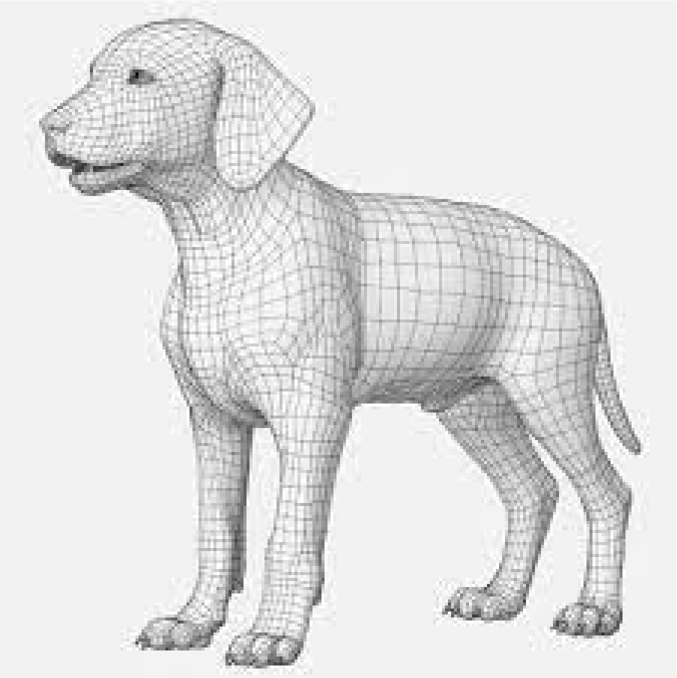
Table 2. Comparison of the detection performance of benchmarking methods with the AUC value at the frame level as the evaluation metric. The **boldface** and the underbar indicate the best and the second-best results, respectively. The *italics* means it does not specify frame or video level AUC. The AUC results of DefakeHop is reported in both frame-level and video-level. The AUC results of benchmarking methods are taken from [19] and [20]. [a] deep learning method, [b] non deep learning method.

		1st Generation datasets		2nd Generation datasets		
	Method	UADFV	FF++ / DF	Celeb-DF v1	Celeb-DF v2	Number of parameters
Zhou *et al.*.(2017) [3]	InceptionV3[a]	85.1%	70.1%	55.7%	53.8%	24M
Afchar *et al.*.(2018) [4]	Meso4[a]	84.3%	84.7%	53.6%	54.8%	27.9K
Li *et al.*.(2018) [17]	FWA[a] (ResNet-50)	97.4%	80.1	53.8%	56.9%	23.8M
Yang *et al.*.(2019) [9]	HeadPose[b] (SVM)	89%	47.3%	54.8%	54.6%	-
Matern *et al.*.(2019) [11]	VA-MLP[b]	70.2%	66.4%	48.8%	55%	-
Rossler *et al.*.(2019) [2]	Xception-raw[a]	80.4%	**99.7%**	38.7%	48.2%	22.8M
Nguyen *et al.*.(2019) [5]	Multi-task[a]	65.8%	76.3%	36.5%	54.3%	-
Nguyen *et al.*.(2019) [6]	CapsuleNet[a]	61.3%	96.6%	-	57.5%	3.9M
Sabir *et al.*.(2019) [8]	*DenseNet+RNN*[a]	-	99.6%	-	-	25.6M
Li *et al.*.(2020) [17]	DSP-FWA[a] (SPPNet)	97.7%	93%	-	64.6%	-
Tolosana *et al.*.(2020) [1]	*Xception*[a]	**100%**	99.4%	83.6%	-	22.8M
Ours	DefakeHop (Frame)	**100%**	95.95%	93.12%	87.65%	42.8K
	DefakeHop (Video)	**100%**	97.45%	**94.95%**	**90.56%**	42.8K

Chen, Hong-Shuo, Mozhdeh Rouhsedaghat, Hamza Ghani, Shuowen Hu, Suya You, and C-C. Jay Kuo. "DefakeHop: A Light-Weight High-Performance Deepfake Detector." In 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1-6. IEEE, 2021.

Cross-Domain Knowledge Structure

- 3D dog model vs. 2D dog image



Conclusion

- ➢ **Yesterday (the first two decades, 1990-2012)**
  - Unsupervised CBIR
- ➢ **Today (the last decade, 2013- Present)**
  - Heavily supervised CBIR
  - DL-based feature learning
  - Metric learning
- ➢ **Tomorrow (the next decade)**
  - Push the envelop of DL
  - Real world applications
  - Towards green machine learning

Q & A

Thank You!