INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
OUTREACH DIVISION

talent sprint

Siamese Networks

Pair-wise constraints (not just the labels)

Recognition Vs Verification

- Recognition: eg. K –Class recognition
  - Given a sample, decide which one of the K it is
- Verification: Given a sample and a class ID,
  - Say “YES” or “NO”
- What makes these two different?
  - Popular in many situation (eg. Biometrics)
    - Face, Fingerprint
  - Intra-class variation could be higher than inter-class
    - People change appearance over time.

Verification: Test/verification time

- Often done with nearest neighbors
- Find “K” nearest neighbors
  - Too much of compute?
- Find distance to the samples from the class under question
  - Find Mean distance
  - Find Min Distance
- Variable Threshold: Subject specific
  - Some people have high variability in “signature”

Verification: Enrolment/Training

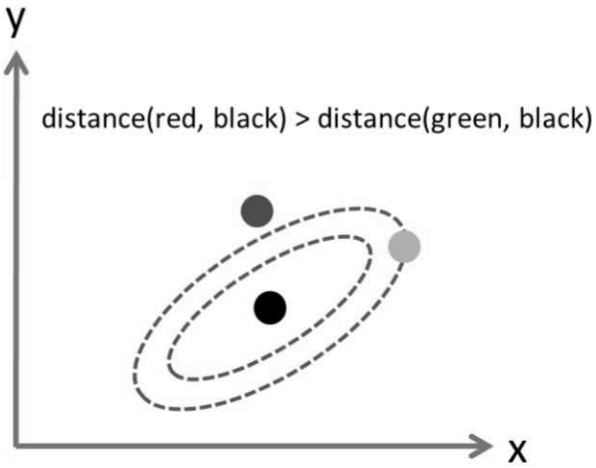
- Usually only small number of samples to add
  - One shot learning (single sample setting)
- Find a feature space that suits the local “geometry”
  - Metric learning
  - Aka finding an appropriate distance function

Mahalanobis distance metric learning

- Mahalanobis distance:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})}.$$

- Mahalanobis distance metric learning



$$d(x, y) = d_A(x, y) = ||x - y||_A = \sqrt{(x - y)^T A (x - y)}$$

$$\begin{aligned} \min_A \quad & \sum_{(x_i, x_j) \in \mathcal{S}} ||x_i - x_j||_A^2 \\ \text{s.t.} \quad & \sum_{(x_i, x_j) \in \mathcal{D}} ||x_i - x_j||_A \geq 1 \\ & A \succeq 0 \end{aligned}$$

Xing E P, Jordan M I, Russell S, et al. Distance metric learning with application to clustering with side-information, NIPS2002.

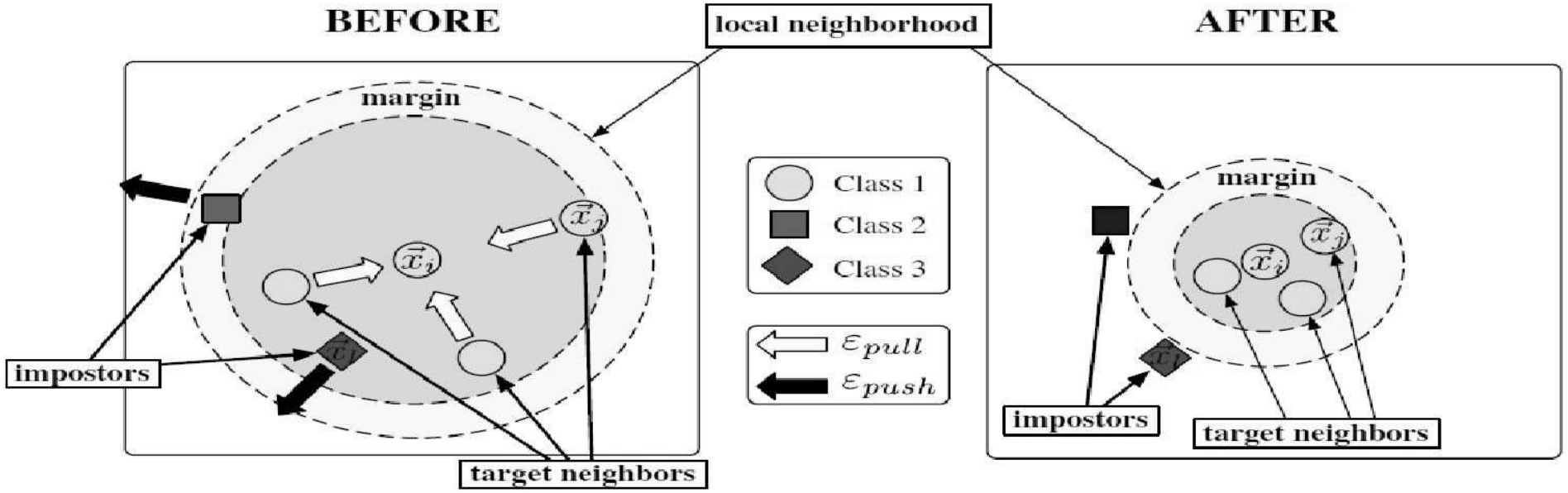
Learning A same as Feature X-form

$$[\mathbf{x}-\mathbf{y}]^T \mathbf{A} [\mathbf{x}-\mathbf{y}] \qquad \mathbf{A} = LL^T$$

Our goal reduced to Finding L or an appropriate feature transformation

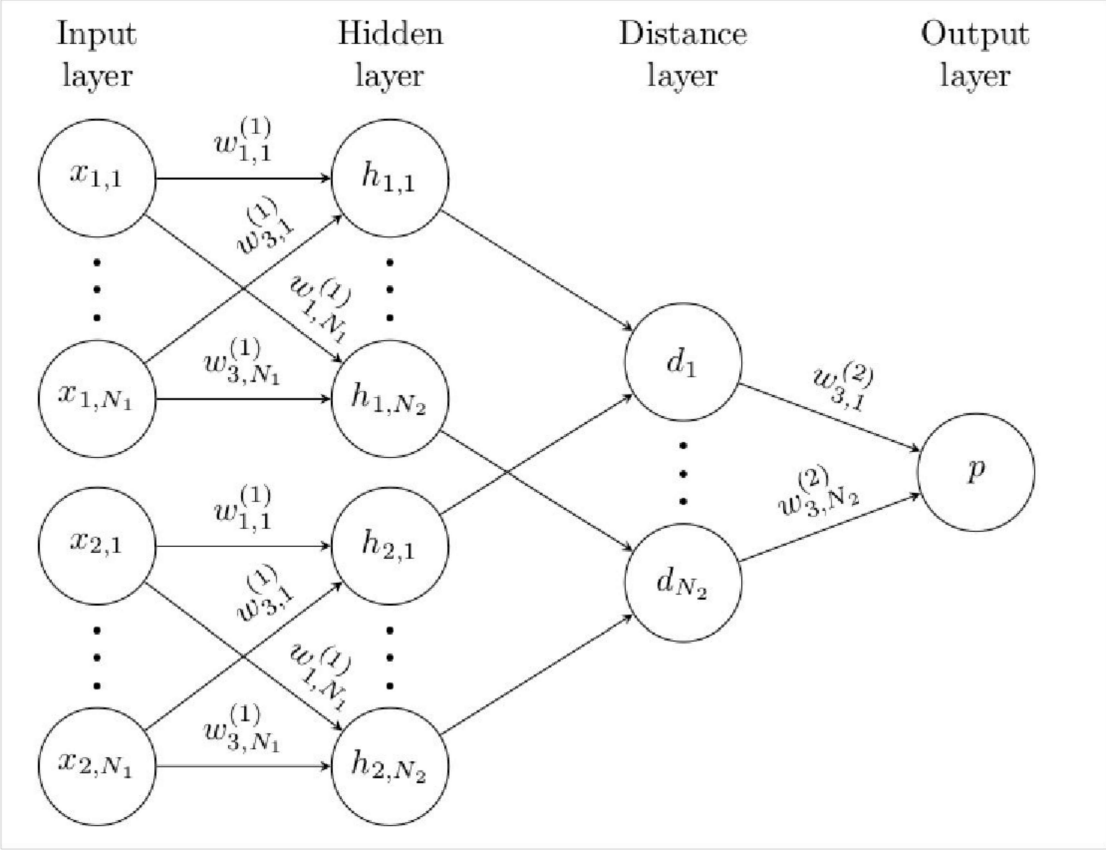
$$[\mathbf{x}-\mathbf{y}]^T LL^T [\mathbf{x}-\mathbf{y}] = [L\mathbf{x}-L\mathbf{y}]^T [L\mathbf{x}-L\mathbf{y}]$$

Large Margin Nearest Neighbour (LMNN)



Weinberger and Saul, JMLR 2009

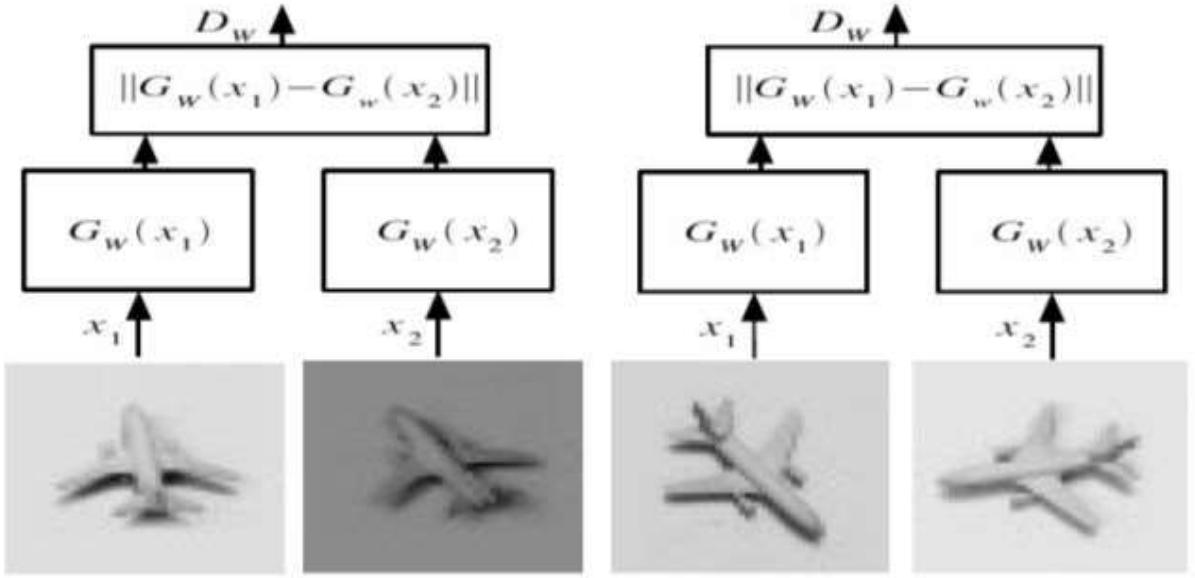
Siamese Nets and Weight Sharing (simple MLP)



Siamese: Loss

Make this small

Make this large



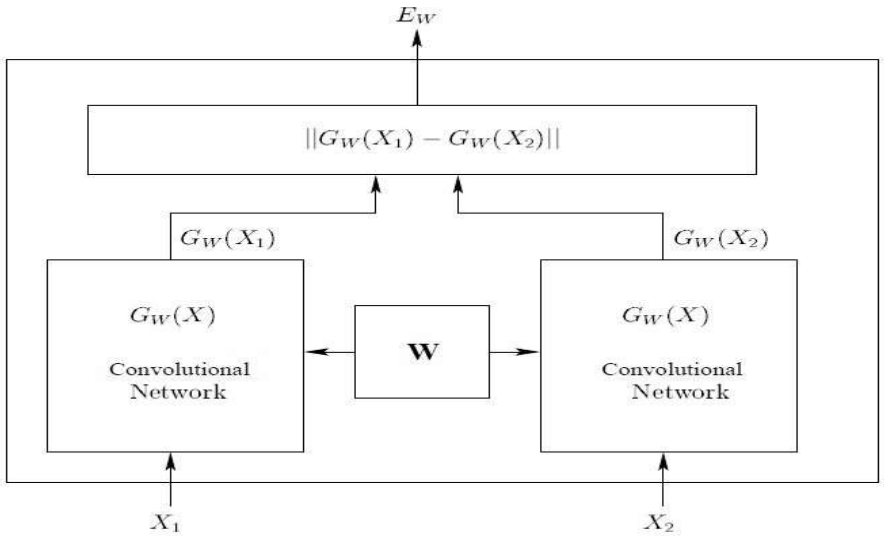
Similar images (neighbors in the neighborhood graph)

Dissimilar images (non-neighbors in the neighborhood graph)

Siamese Architecture

- Given a family of functions $G_W(X)$ parameterized by W, find W such that the similarity metric $D_W(X_1, X_2)$ is small for similar pairs and large for dissimilar pairs:-

$$D_W(X_1, X_2) = ||G_W(X_1) - G_W(X_2)||$$



Loss function

$$\mathcal{L}(W) = \sum_{i=1}^{p} L(W, (Y, \vec{X}_1 \vec{X}_2)^i)$$

Loss function for similar pairs

$$L(W, (Y, \vec{X}_1, \vec{X}_2)^i) = (1 - Y) L_s(D_W^i) + Y L_D(D_W^i)$$

Loss function for dissimilar pairs

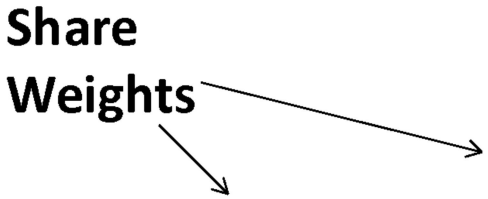
Raia Hadsell, Sumit Chopra, Yann LeCun, Dimensionality Reduction by Learning an Invariant Mapping. CVPR 2006

Contrastive loss function

- Given $D_W(X_1, X_2)$ is the distance between the pair of samples, the contrastive loss function is defined as follows:

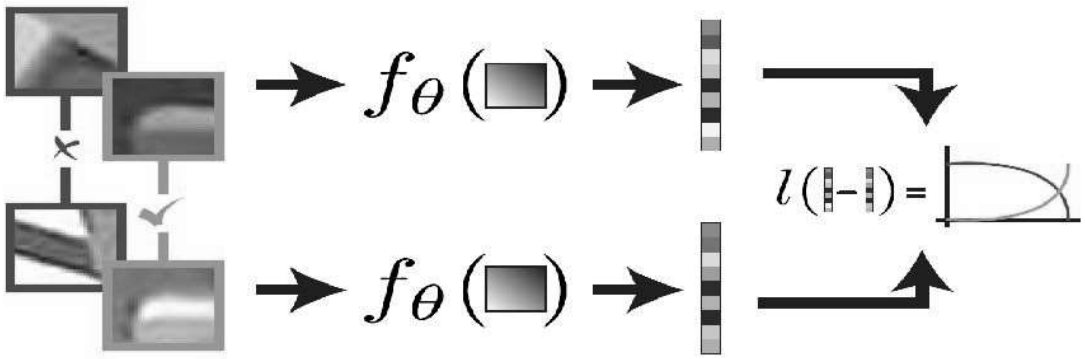
$$(1-Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\{max(0, m - D_W)\}^2$$

Historic



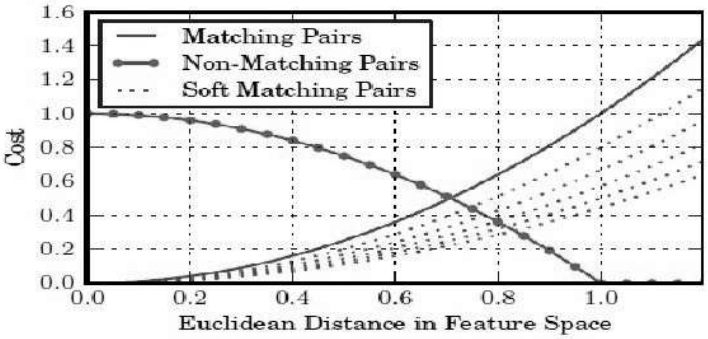
Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E. and Shah, R., 1993. Signature Verification Using A "Siamese" Time Delay Neural Network. *IJPRAI*, *7*(4), pp.669-688.

Application: Learning to Match Siamese Network



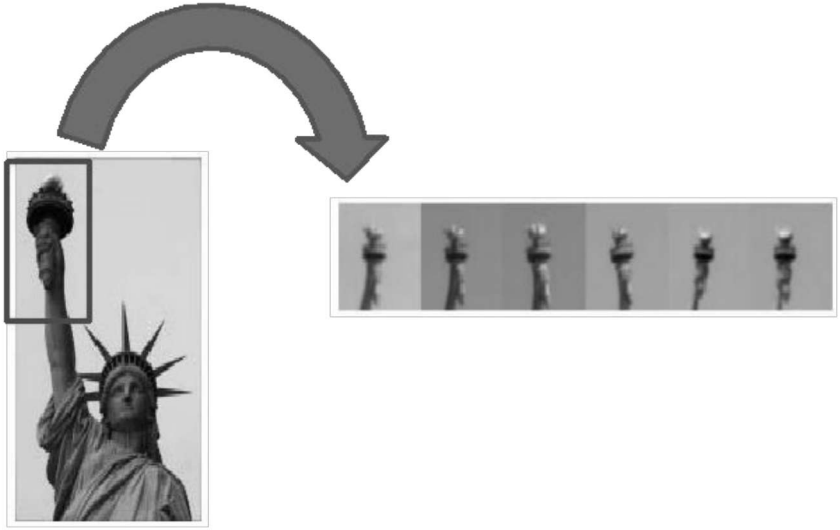
Using the contrastive cost function

$$l_{\boldsymbol{\theta}}\left(\mathbf{y}_i, \mathbf{y}_j\right) = \begin{cases} s_{ij} d_{ij}^2, & \text{if matching} \\ \max\left(1.0 - d_{ij}^2,\ 0\right), & \text{if non-matching} \end{cases}$$



IROS 2014

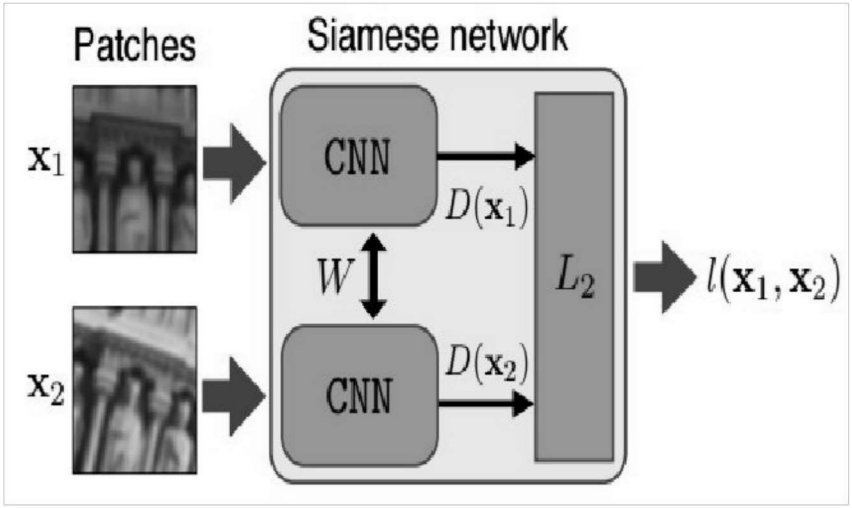
Matching across views



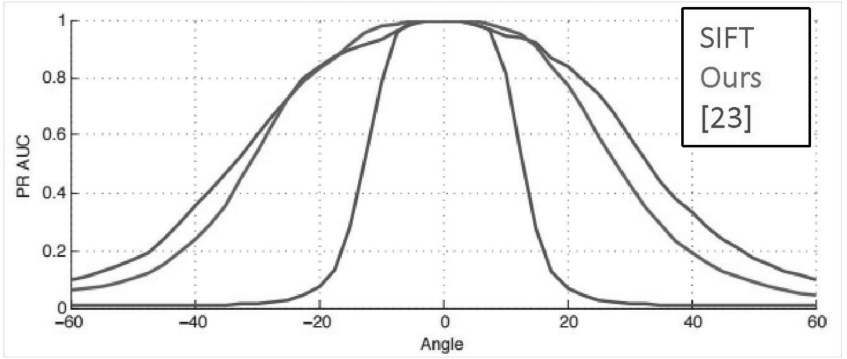
Learn a discriminative representation of patches from different views of 3D points

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P. and Moreno-Noguer, F., 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 118-126).

CNNs to output descriptors (aka “Deep-SIFT”)



Robustness to Rotation



Use the CNN outputs of our Siamese networks as descriptor

Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P. and Moreno-Noguer, F., 2015. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 118-126).

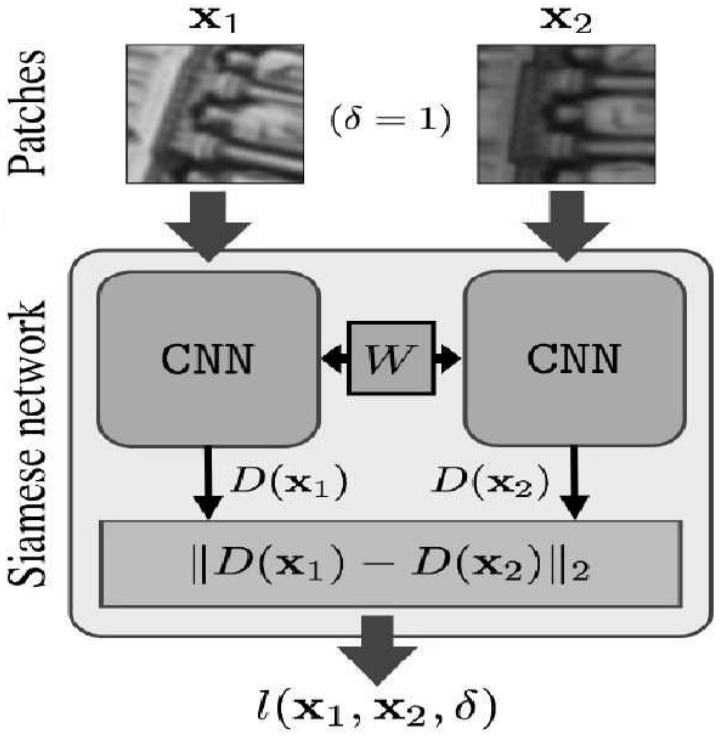
Learning to Match

$$d_D(\mathbf{x}_1, \mathbf{x}_2) = \|D(\mathbf{x}_1) - D(\mathbf{x}_2)\|_2$$

$$l(\mathbf{x}_1, \mathbf{x}_2, \delta) = \delta \cdot l_P(d_D(\mathbf{x}_1, \mathbf{x}_2)) + (1 - \delta) \cdot l_N(d_D(\mathbf{x}_1, \mathbf{x}_2))$$

$$l_P(d_D(\mathbf{x}_1, \mathbf{x}_2)) = d_D(\mathbf{x}_1, \mathbf{x}_2)$$

$$l_N(d_D(\mathbf{x}_1, \mathbf{x}_2)) = \max(0, m - d_D(\mathbf{x}_1, \mathbf{x}_2))$$



Stochastic sampling and aggressive mining

ICLR 2014

Person reidentification

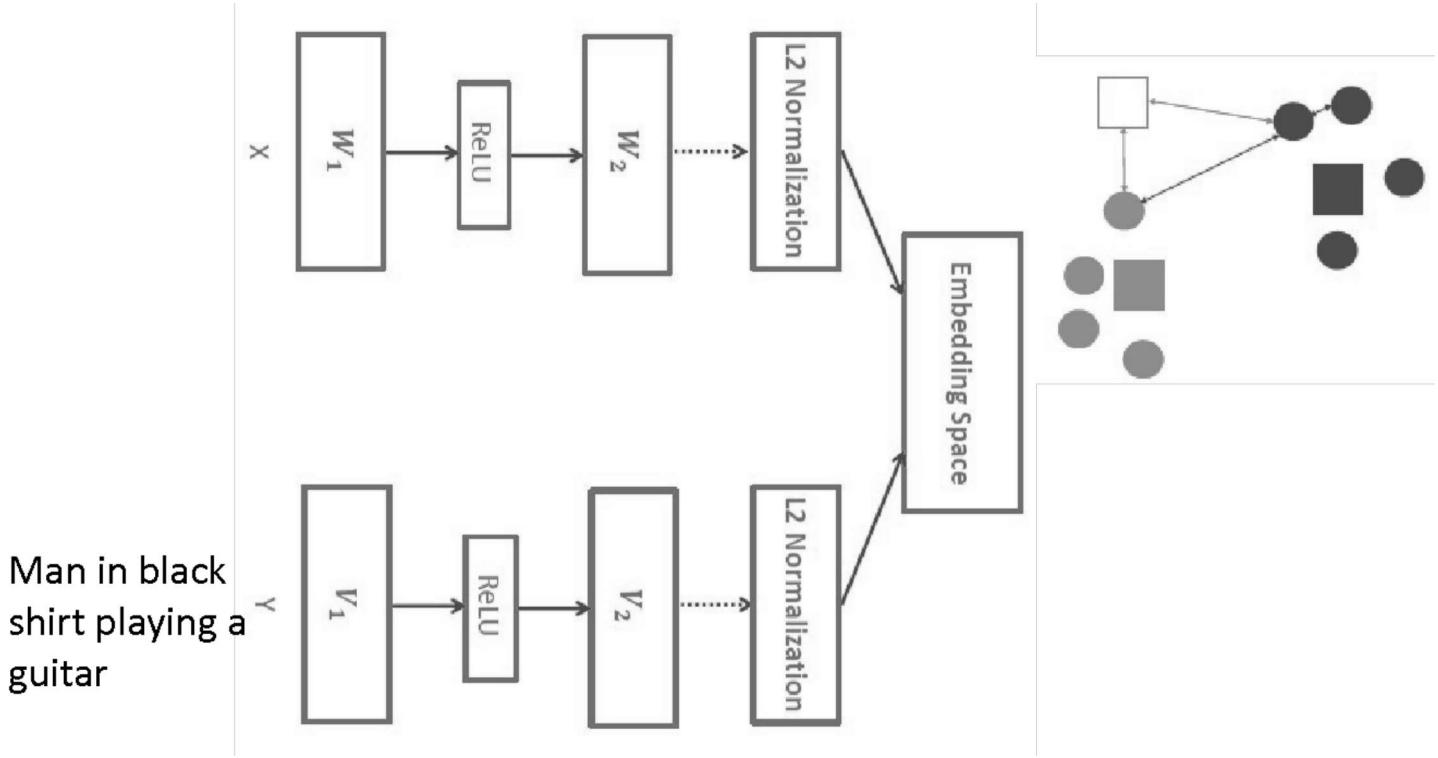
CUHK03 Data set

True positive

True negative

Ahmed, E., Jones, M. and Marks, T.K., 2015. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3908-3916).

Cross modal matching



Wang, L., Li, Y. and Lazebnik, S., 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5005-5013).

Another application

Application: Sentence completion, response to tweet, paraphrase identification

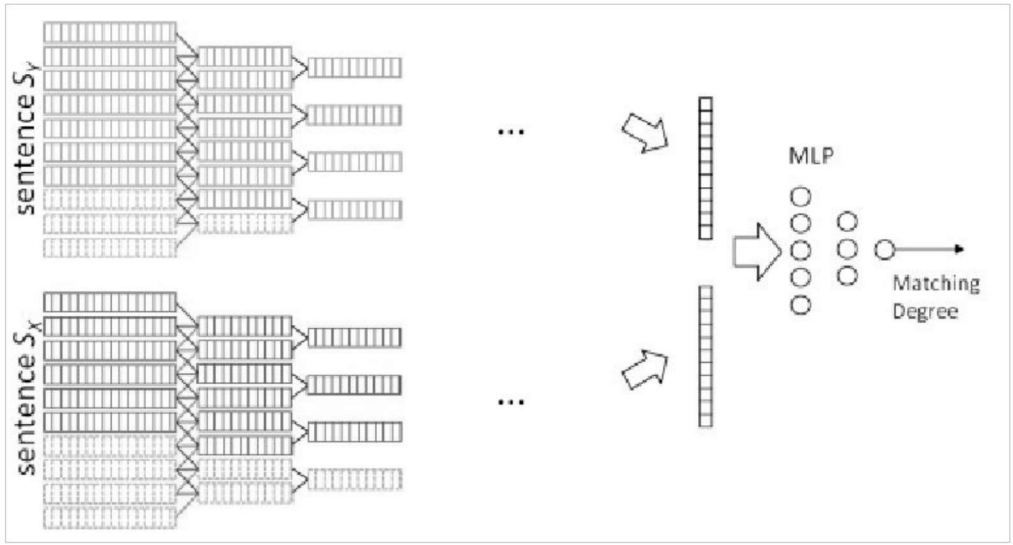
Example:

x : Damn, I have to work overtime this weekend!

y^+: Try to have some rest buddy.

y^-: It is hard to find a job, better start polishing your resume.

word2vec



Hu, Baotian, et al., Convolutional neural network architectures for matching natural language sentences, NIPS 2014

Triplet Loss

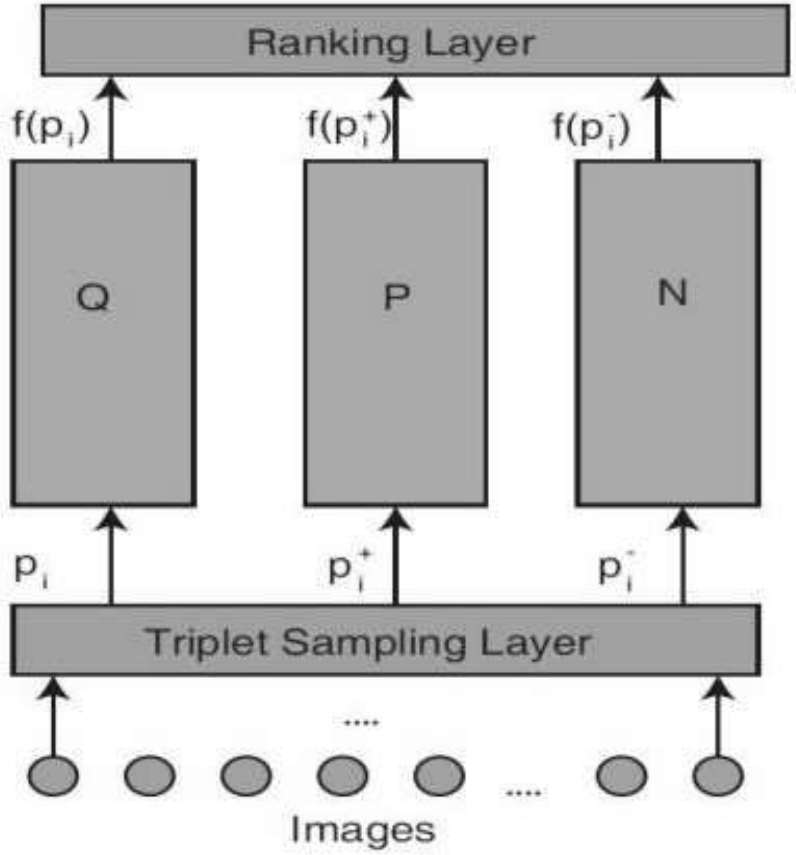


Image Retrieval (Ranking)

Summary

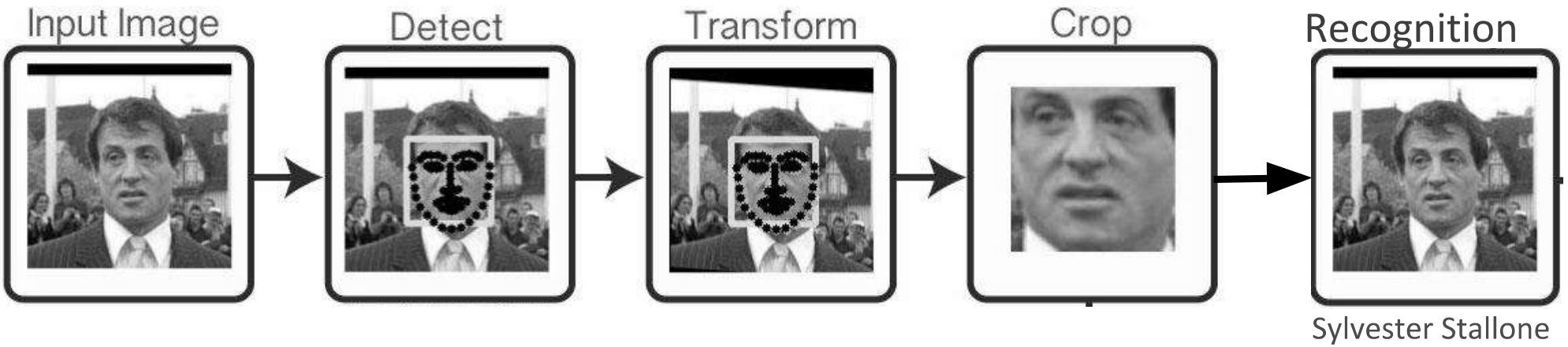
- Siamese and Triplet networks/losses are popular for solving fine grain classification (e.g. Face), capturing subjective needs (eg. Invariance to rotation), rankings etc.

- Available in many frameworks.

- Number of examples increase (e.g. ${}^{n}C_{2}$, ${}^{n}C_{3}$). Finding right pairs/triplets is also important.

Focus: Face

Face processing pipeline

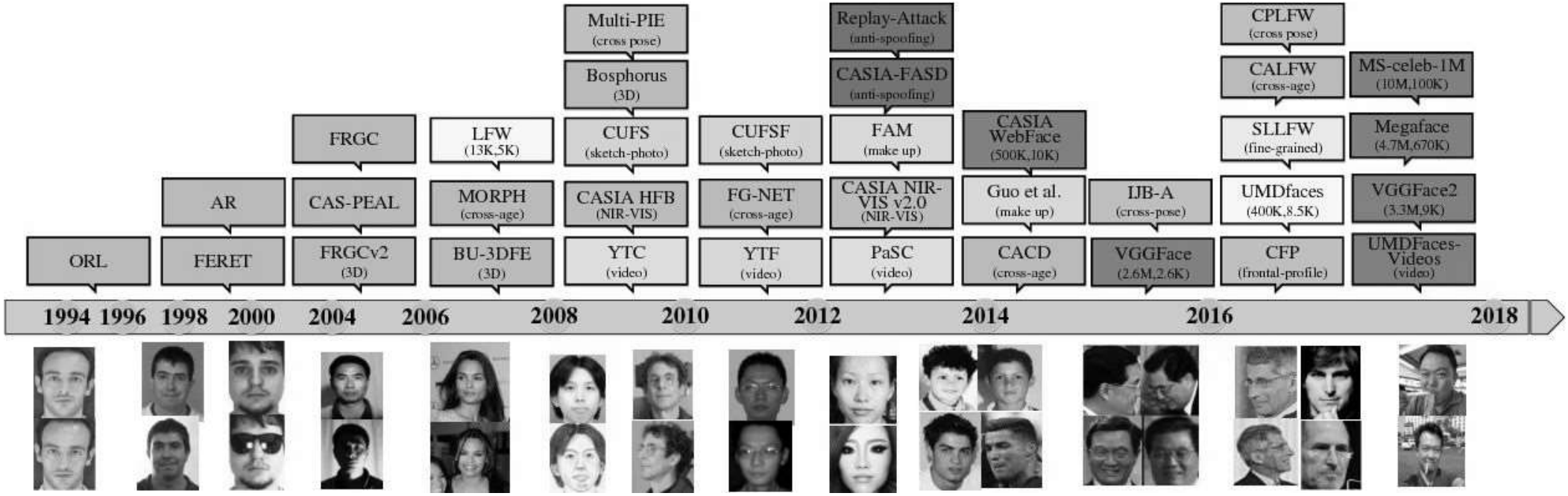


Evolution of FR Datasets

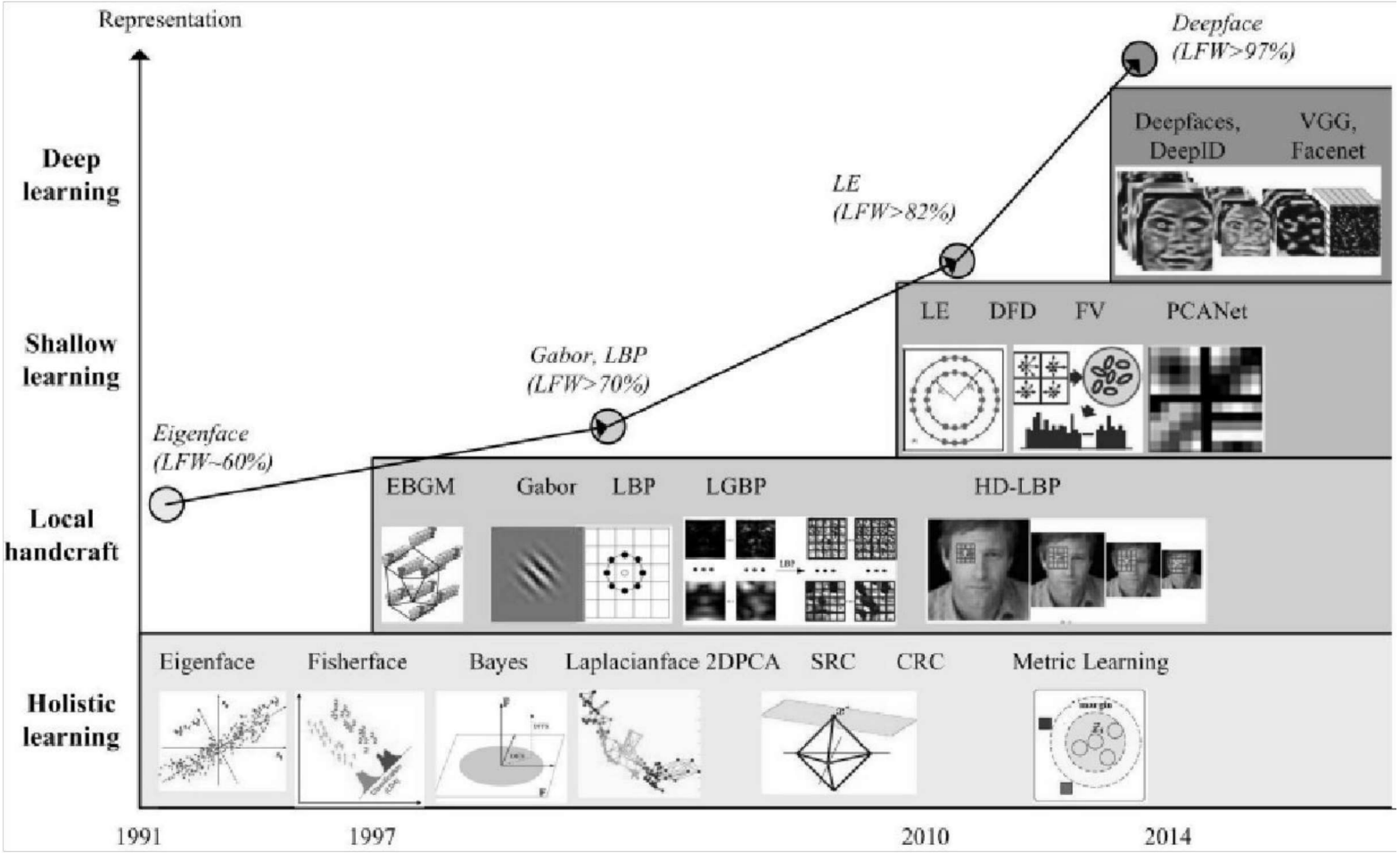
Controlled and Small

Emergence of Wild

Large and Suitable for Training Deep



Performance on LFW



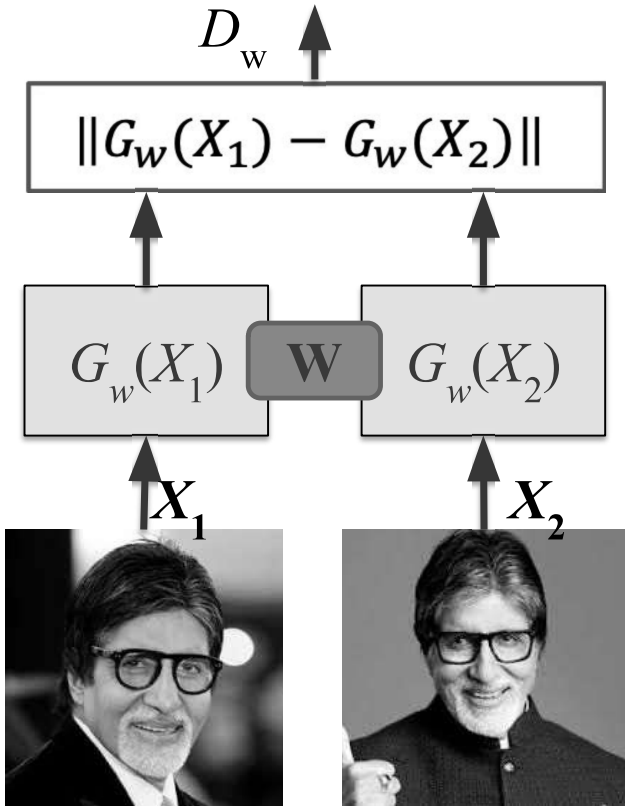
Systematic Evolution of Performance (LFW)

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [160]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [152]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [153]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [144]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [105]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [123]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [188]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [181]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [107]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [224]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [129]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [171]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [111]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [62]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [39]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [106]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [128]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [170]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [172]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [38]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [235]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

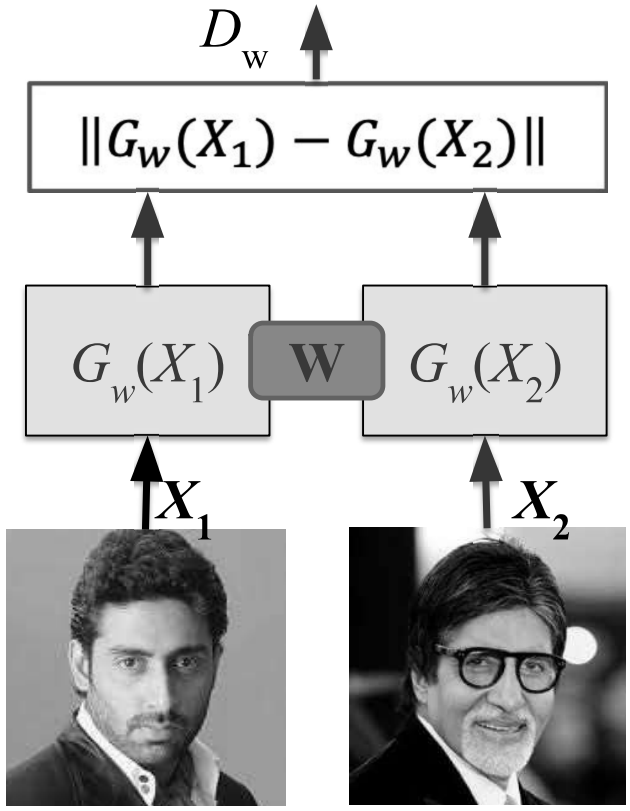
M. Wang and W. Deng, “Deep Face Recognition: A Survey”, Arxiv, 2018

Siamese Loss

Make this smaller

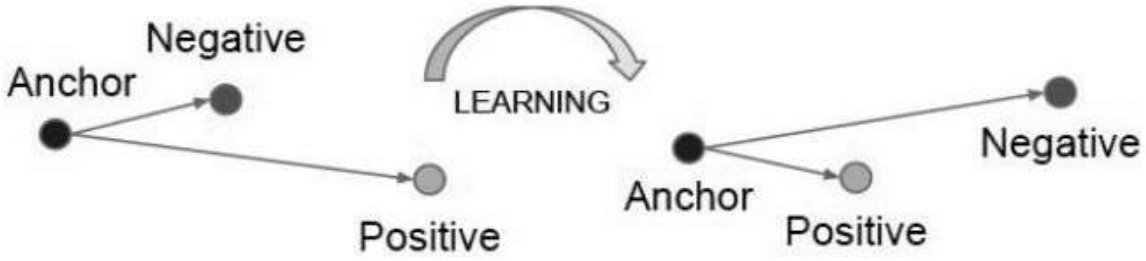


Make this larger



- Only pair-wise Labels
- Similarity Metric: $D_w(X_1, X_2)$
- Have shared weights
- Training in batches

Triplet loss function



$$\|x_i^a - x_i^p\|_2^2 + \alpha < \|x_i^a - x_i^n\|_2^2, \forall (x_i^a, x_i^p, x_i^n) \in \mathcal{T}$$

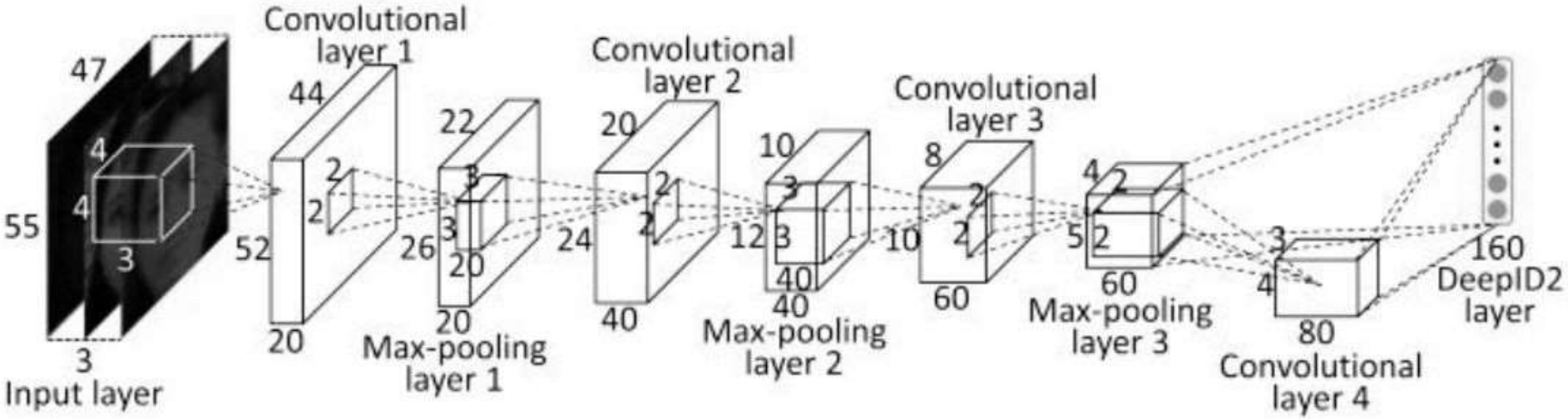
- Triplet loss function:

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

where *f* is the embedding

Florian Schroff et al. FaceNet: A Unified Embedding for Face Recognition and Clustering, CVPR 2015

Deep ID2



Yi Sun, etc. Deep Learning Face Representation by Joint Identification-Verification. NIPS 2014

Deep ID2: Identification + Verification

- Verification loss:

$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} \frac{1}{2}\|f_i - f_j\|_2^2 & \text{if } y_{ij} = 1 \\ \frac{1}{2}\max\left(0, m - \|f_i - f_j\|_2\right)^2 & \text{if } y_{ij} = -1 \end{cases}$$

- Identity loss:

$$\text{Ident}(f, t, \theta_{id}) = -\sum_{i=1}^{n} -p_i \log \hat{p}_i = -\log \hat{p}_t$$

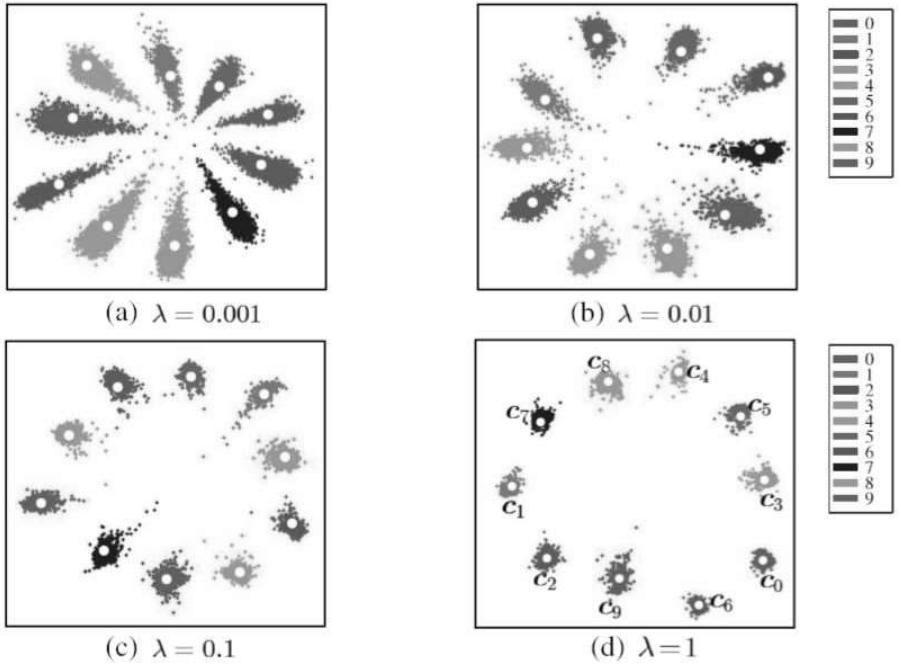
- Total loss = Identity_Loss + λVerification_Loss

Center loss

Center loss: $\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$ Where c is the center of the cluster

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

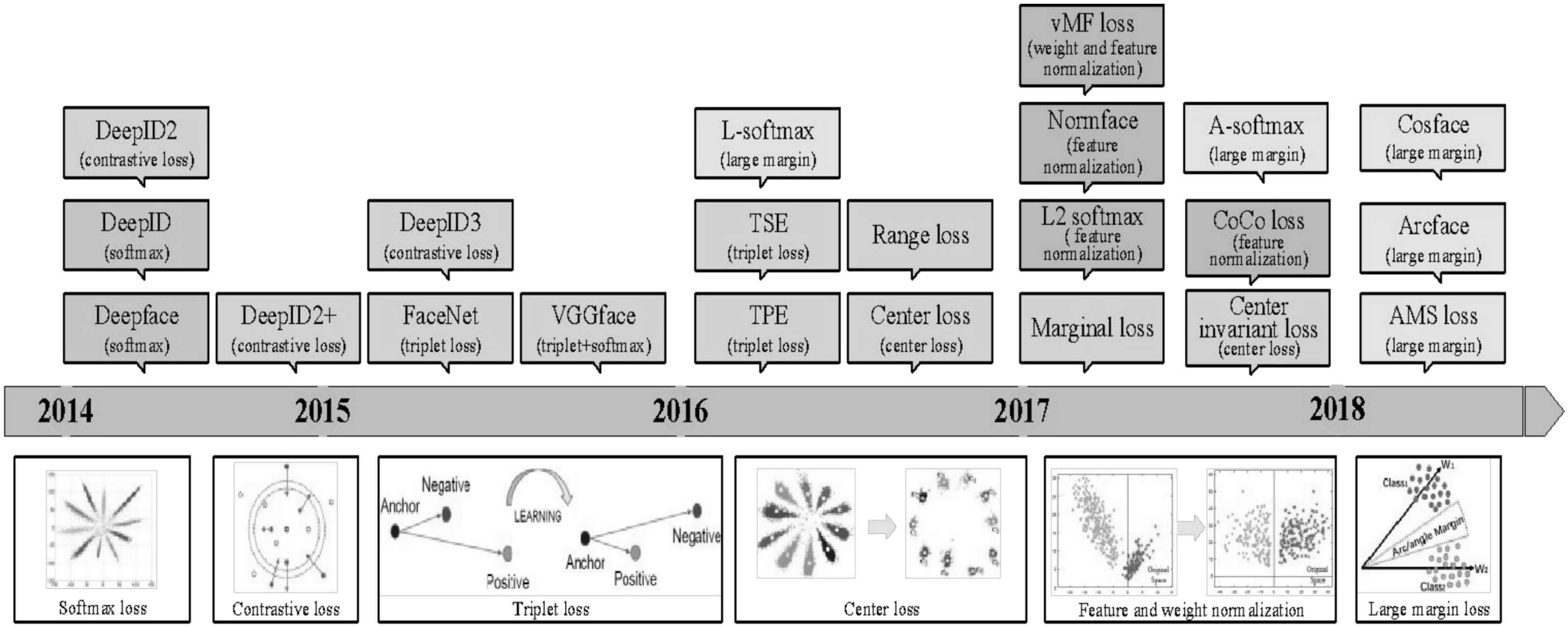
$$= -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T \boldsymbol{x}_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$



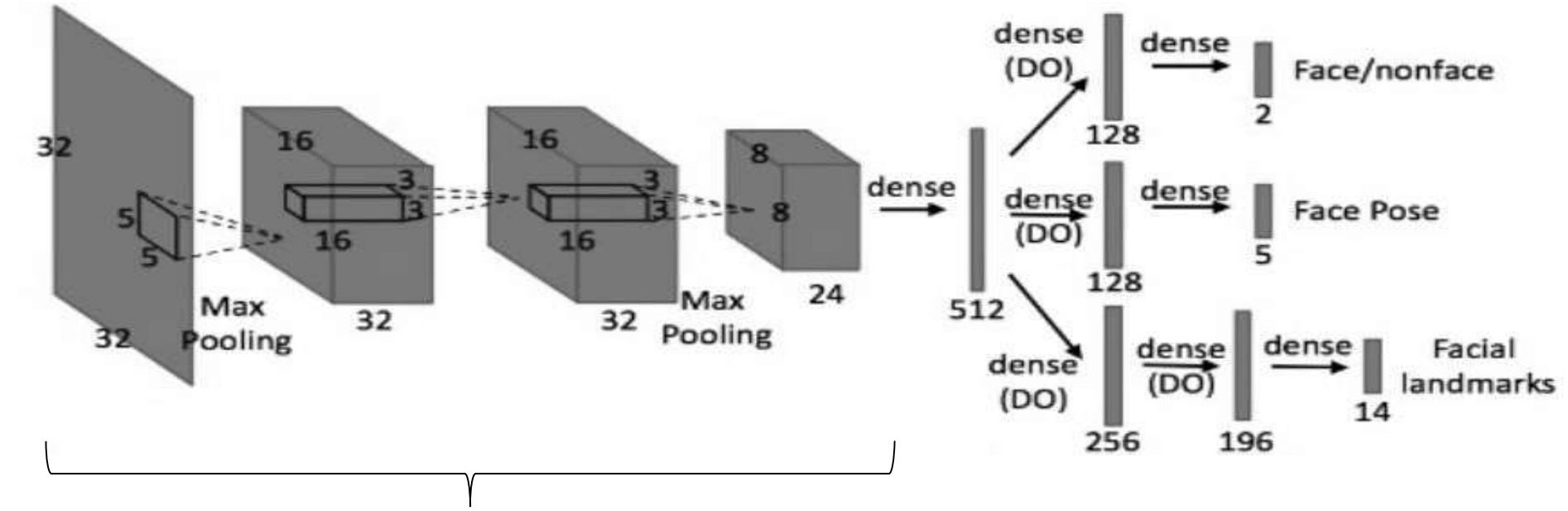
(a) $\lambda = 0.001$ (b) $\lambda = 0.01$

(c) $\lambda = 0.1$ (d) $\lambda = 1$

Summary: Evolution of Loss Functions



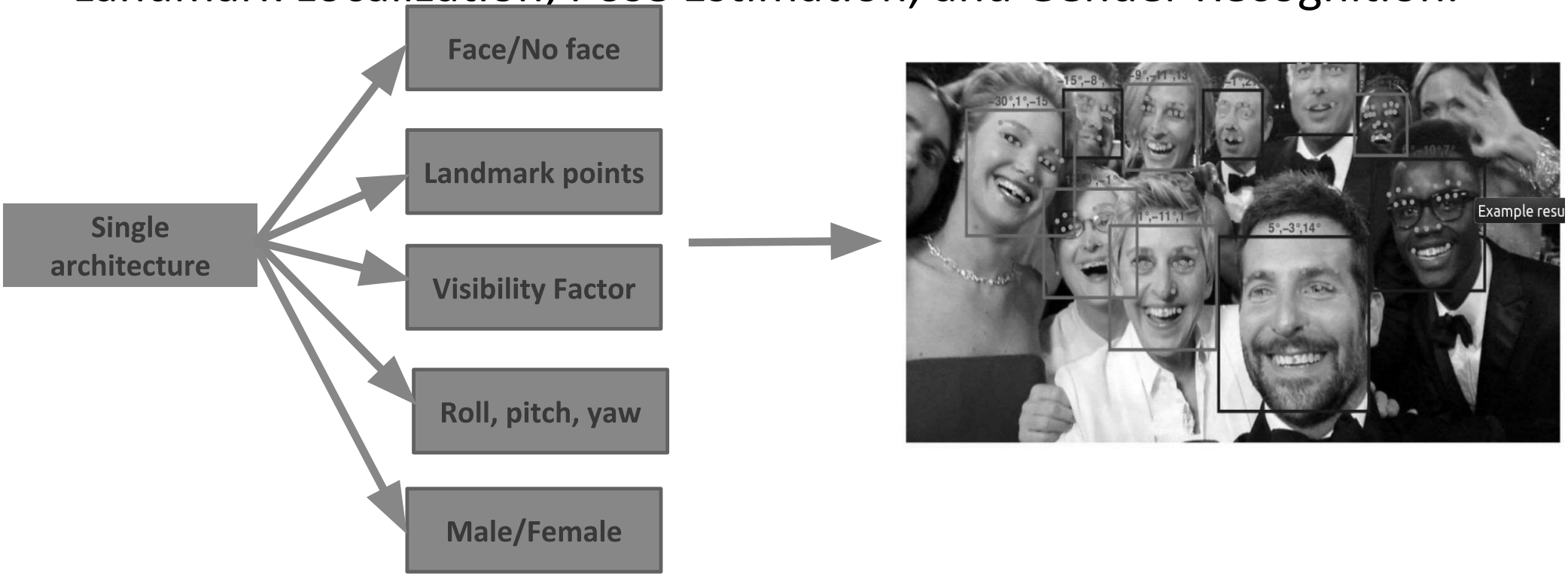
Multi task network?



Shared hidden layers
Shared representation

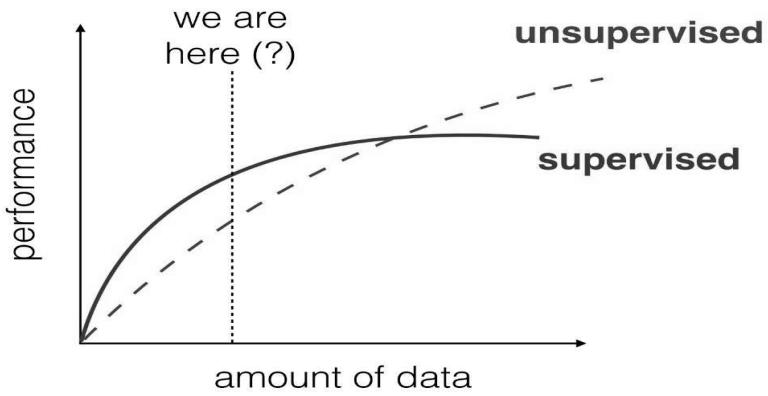
Hyperface: state of art

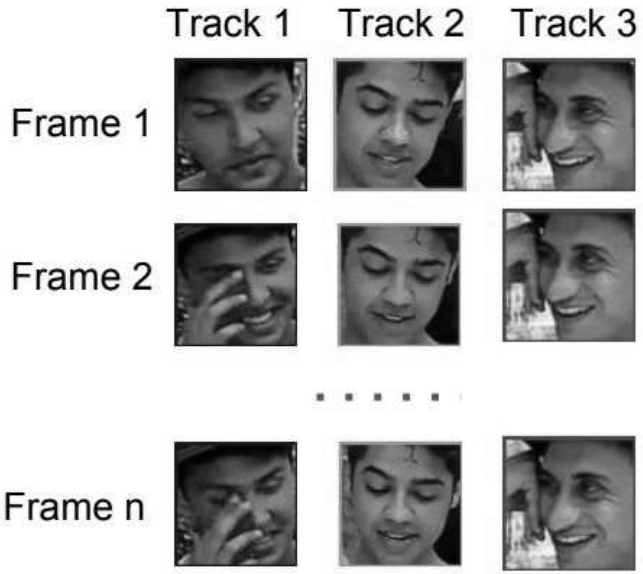
- A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition.



Unsupervised Learning of Face Representations

Samyak Datta, Gaurav Sharma and C. V. Jawahar,
IEEE FG, 2018 (Best Paper Award)

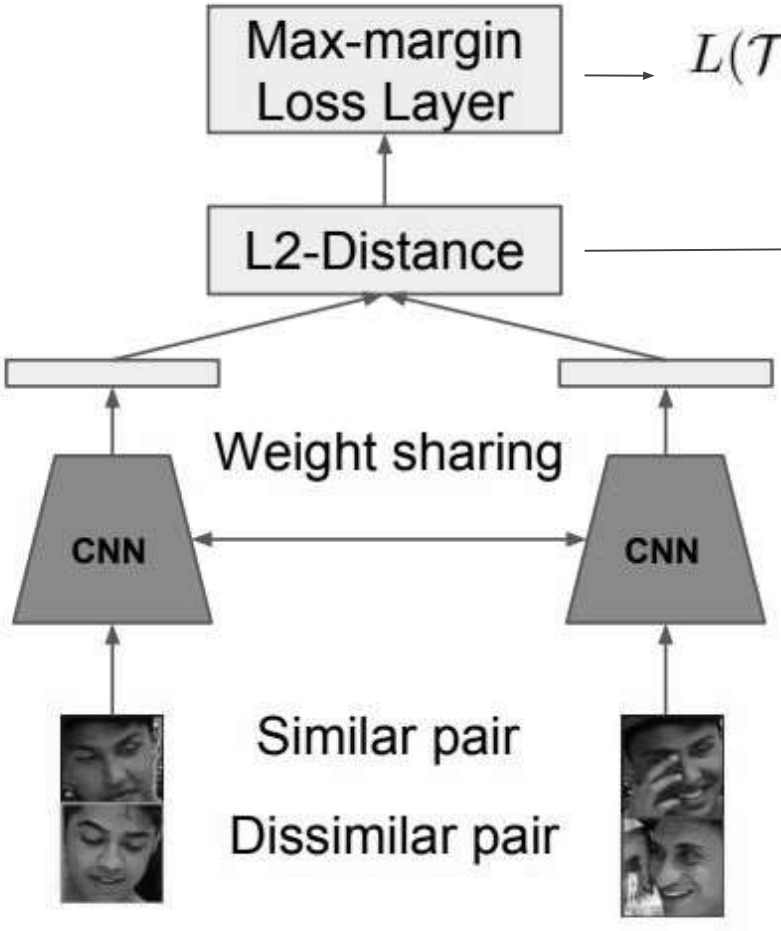




Learn from YouTube

1. Multiple faces in the same frame must belong to different people
2. The same face tracked across multiple frames belongs to the same identity

Model



$$L(\mathcal{T}) = \sum_{\mathcal{T}} \max\left\{0, m - y_{ij}(b - D^2(\mathbf{x}_i, \mathbf{x}_j))\right\}$$

$$D^2(\mathbf{x_1}, \mathbf{x_2}) = \|\phi(\mathbf{x_1}) - \phi(\mathbf{x_2})\|_2^2$$

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD
OUTREACH DIVISION

talent sprint

Thanks!!

Questions?