CSE 252D: Advanced Computer Vision

Manmohan Chandraker

Lecture 10: Face Recognition 2

Virtual classrooms

- Virtual lectures on Zoom
  - Only host shares the screen
  - Keep video off and microphone muted
  - But please do speak up (remember to unmute!)
  - Slides uploaded on webpage just before class
- Virtual interactions on Zoom
  - Ask and answer plenty of questions
  - “Raise hand” feature on Zoom when you wish to speak
  - Post questions on chat window
  - Happy to try other suggestions!
- Lectures recorded and upload on Canvas
  - Available under “My Media” on Canvas

Overall goals for the course

- Introduce the state-of-the-art in computer vision
- Study principles that make them possible
- Get understanding of tools that drive computer vision
- Enable one or all of several such outcomes
  - Pursue higher studies in computer vision
  - Join industry to do cutting-edge work in computer vision
  - Gain appreciation of modern computer vision technologies
- This is a great time to study computer vision!

Papers for Wed, May 05

- A Discriminative Feature Learning Approach for Deep Face Recognition
  - https://ydwen.github.io/papers/WenECCV16.pdf
- SphereFace: Deep Hypersphere Embedding for Face Recognition
  - https://arxiv.org/abs/1704.08063
- ArcFace: Additive Angular Margin Loss for Deep Face Recognition
  - https://arxiv.org/abs/1801.07698
- CosFace: Large Margin Cosine Loss for Deep Face Recognition
  - https://arxiv.org/abs/1801.09414

Papers for Fri, May 07

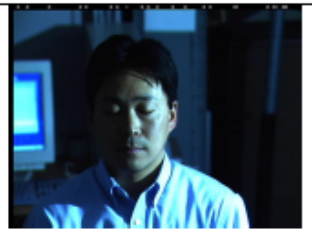
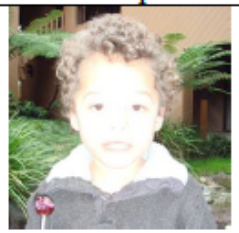
- Deep High-Resolution Representation Learning for Human Pose Estimation
  - https://arxiv.org/abs/1902.09212
- Simple Baselines for Human Pose Estimation and Tracking
  - https://arxiv.org/abs/1804.06208
- OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields
  - https://arxiv.org/abs/1812.08008
- End-to-end Recovery of Human Shape and Pose
  - https://arxiv.org/abs/1712.06584

Papers for Wed, May 12

- ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation
  - https://arxiv.org/abs/1606.02147
- ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation
  - https://ieeexplore.ieee.org/abstract/document/8063438
- Fast-SCNN: Fast Semantic Segmentation Network
  - https://arxiv.org/abs/1902.04502
- Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation
  - https://arxiv.org/abs/1506.04924

Recap

Unconstrained Face Recognition

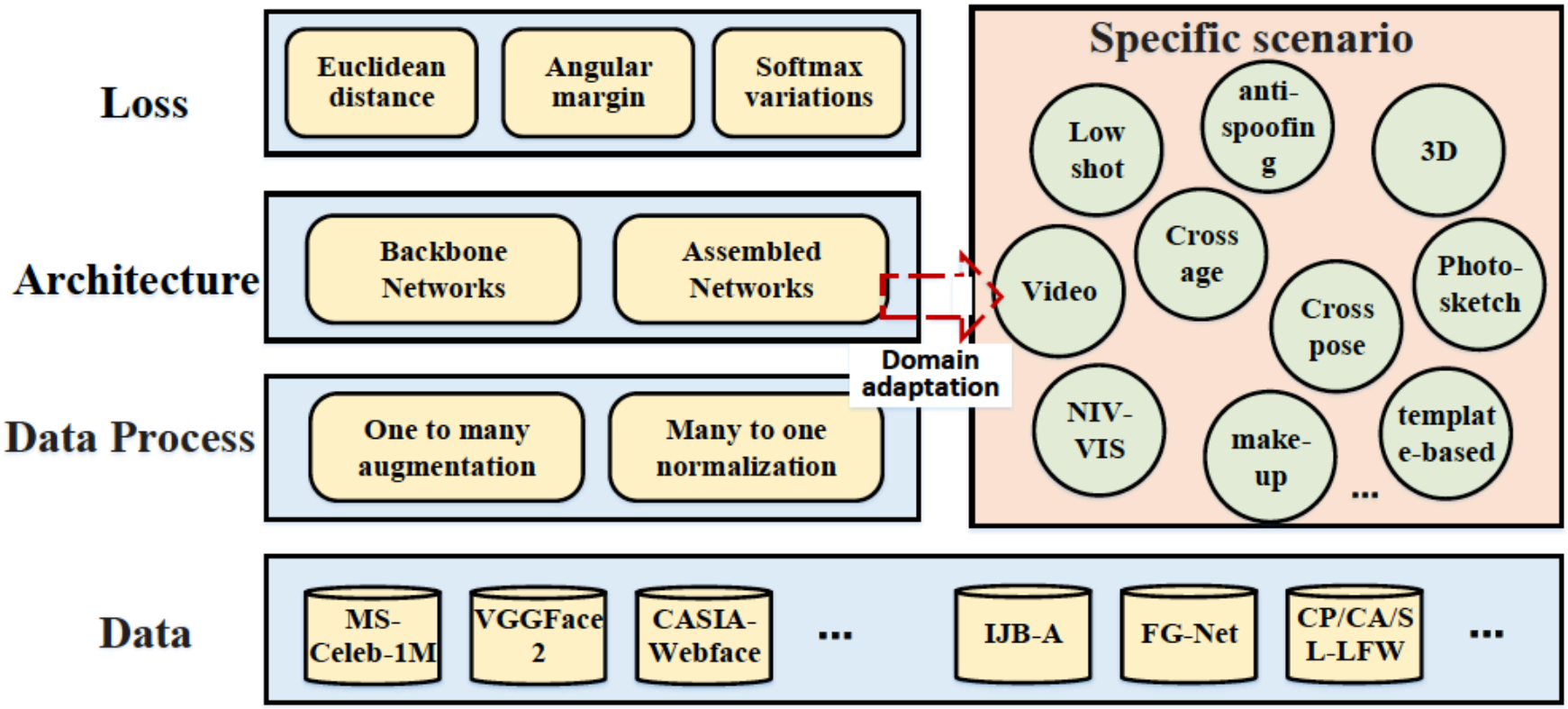
Scenario	External occlusion	Self occlusion	Facial accessories	Limited field of view (FOV)	Extreme illumination	Sensor saturation
Examples	occlusion by other objects	non-frontal pose	hat, sunglasses, scarf, mask	partially out of camera's FOV	gloomy or highlighted facial area	underexposure or overexposure
Image						

[Liao et al., Partial Face Recognition, PAMI 2013]

Face Recognition on LFW Benchmark

- Human performance : **99.20%**
- Local Binary Patterns : 95.17%
- DeepFace : 97.35 %
- DeepID2 : 99.15%
- FaceNet : **99.63%**

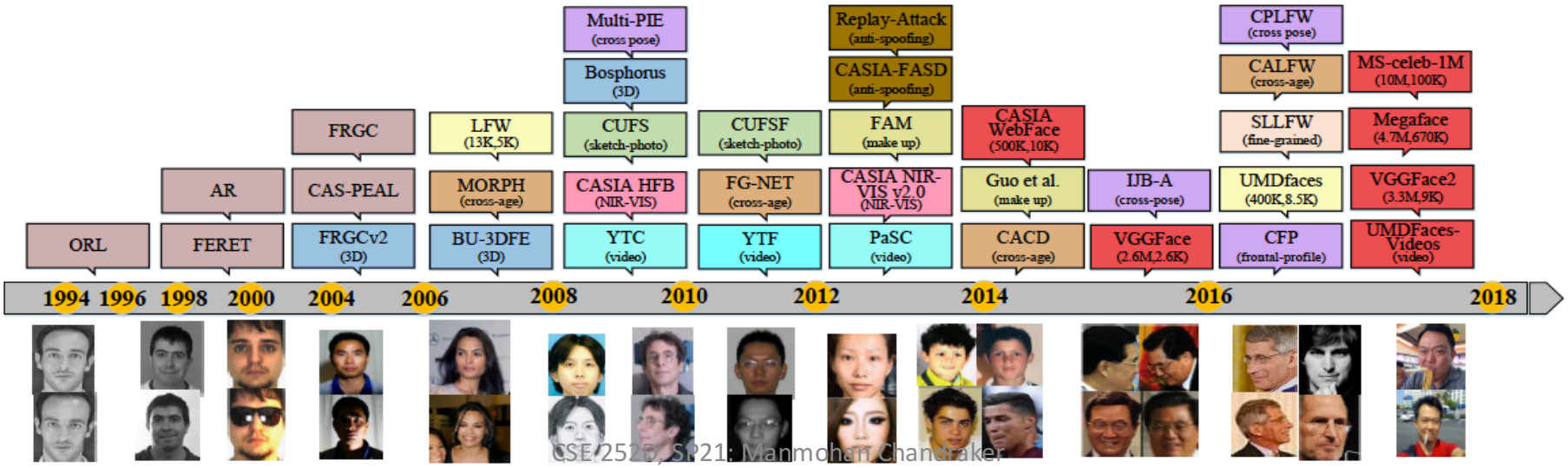
Axes for Studying Face Recognition



[Wang and Deng, "Deep Face Recognition: A Survey"]

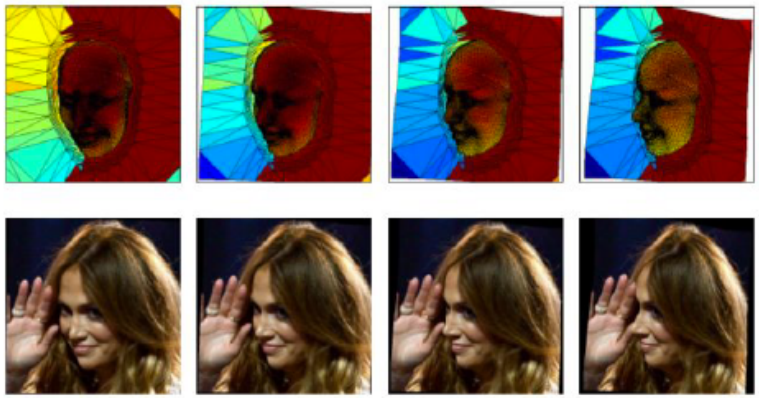
Face Datasets

Name	Identities	Images	Purpose
LFW	5,749	13,233	Small, used for testing, saturated
Celeb Faces	10177	202,599	Many identities, attribute labels
VGG-Face	2622	1,635,159	Many examples per class, somewhat noisy
IJB-A	500	5k images, 2k videos	Challenging pose, lighting, quality
MS-1M	80k	7M	Largest public dataset for training (currently)
Facebook	4030	4.4M	Proprietary, used in DeepFace
Google	8M	200M	Proprietary, used in FaceNet



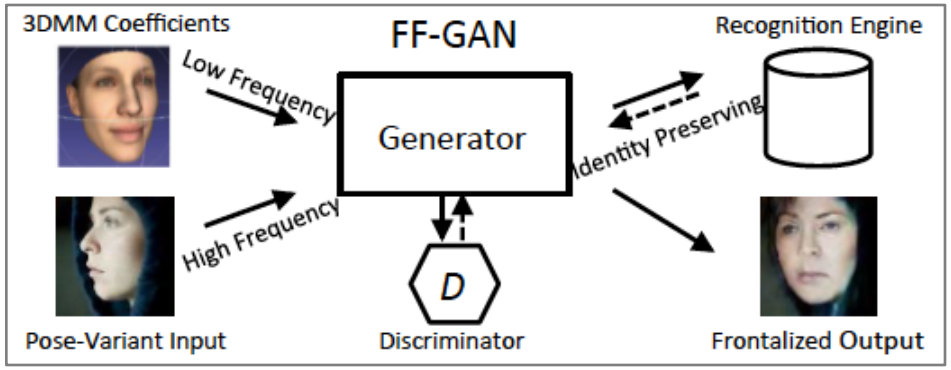
Face Processing

Augmentation



[Zhu et al., "3DDFA"]

Normalization



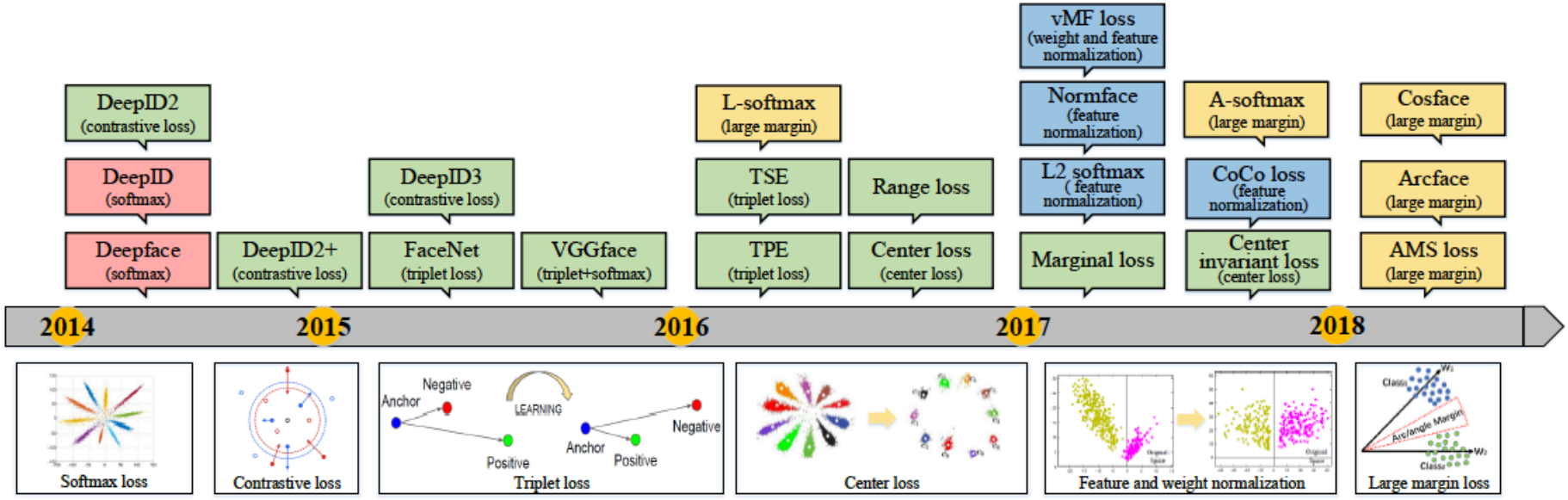
[Yin et al., "FF-GAN"]

- **One-to-Many Augmentation:** mitigate difficulty of diverse data collection
  - Generate 3D pose-variant faces from frontal inputs, use for training
  - Use GANs or other methods to generate faces with diverse attributes

- **Many-to-One Normalization:** reduce variation in test-time inputs
  - Generate frontal face from pose-variant input
  - Use GANs or methods to generate faces with neutral attributes

Loss Functions

Large-margin losses and softmax variants



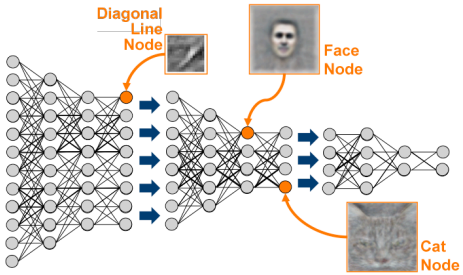
Learning Face Representations

Steps in Face Recognition

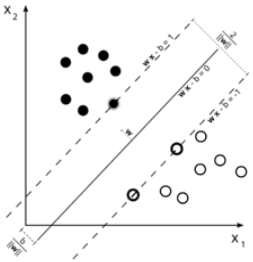
- Face Detection
  - Localize the face

- Face Alignment
  - Factor out 3D transformation

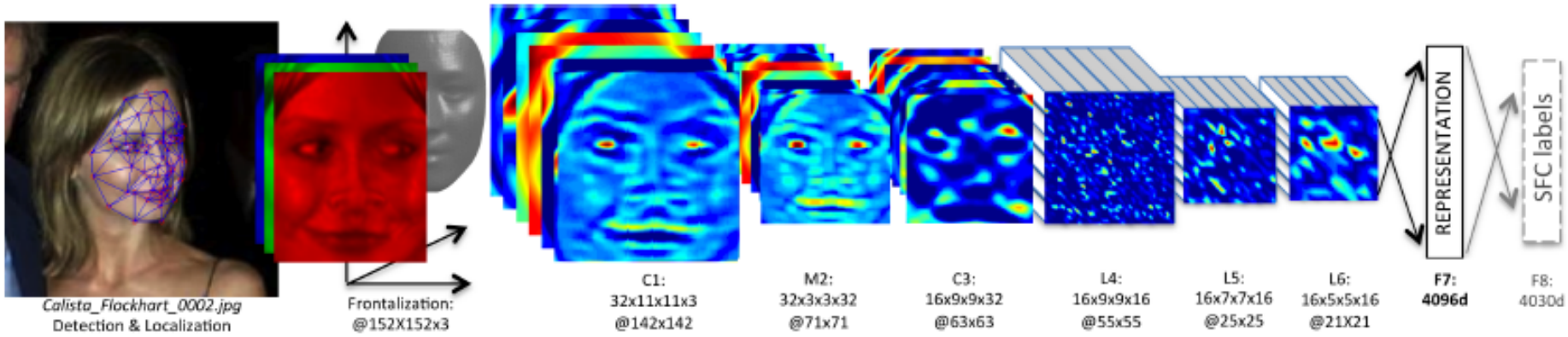
- Feature Extraction
  - Find compact representation



- Classification
  - Answer the question



Architecture



Layer 1-3 : Intuition

- Convolution layers - extract low-level features (e.g. simple edges and texture)

Layer 4-6: Intuition

- Apply filters to different locations on the map
- Similar to a conv. layer but spatially dependent

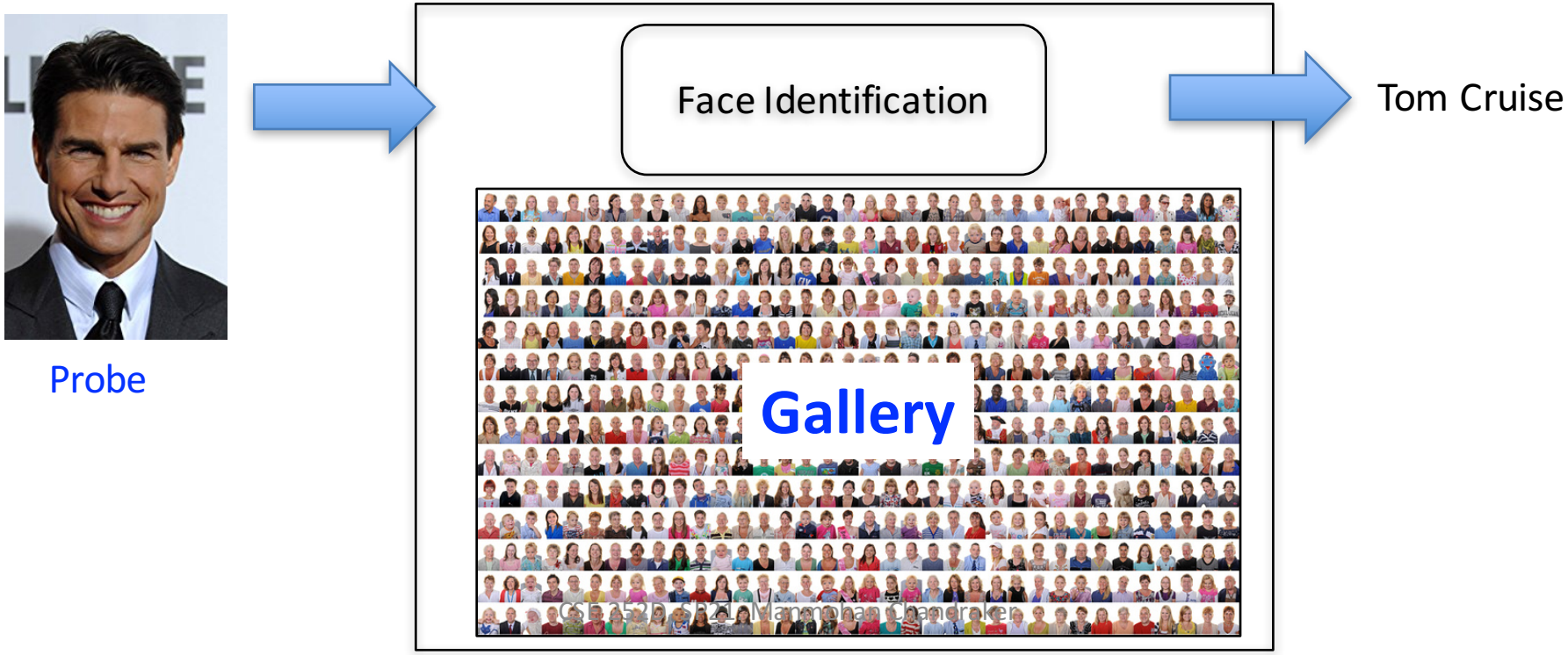
- Different regions of an aligned image have different local statistics
  - Aligned images with similar semantic concepts are being considered
  - A large training dataset is available, can handle increased parameters

[Taigman et al., DeepFace, CVPR 2014]

Face Identification

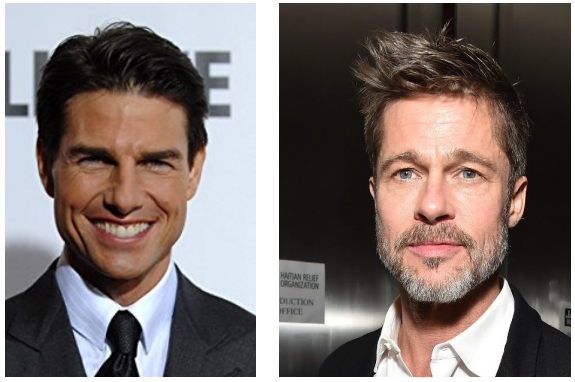
- Closed set identification: assign one of gallery identities to probe image
- Galleries can be very large, high chance of similar appearances

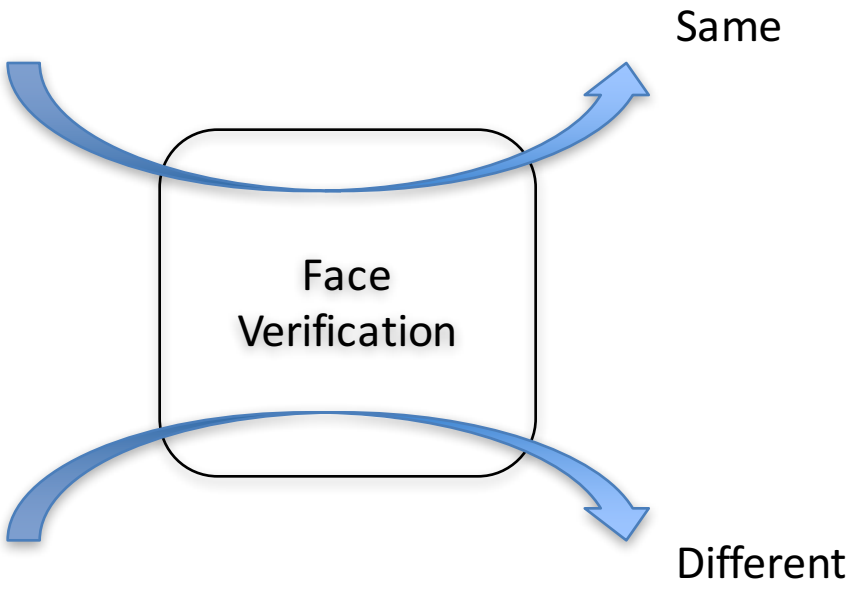
- Goal is to have sharp decision boundary between gallery identities
- Feature need not generalize to other tasks (identities outside the gallery)



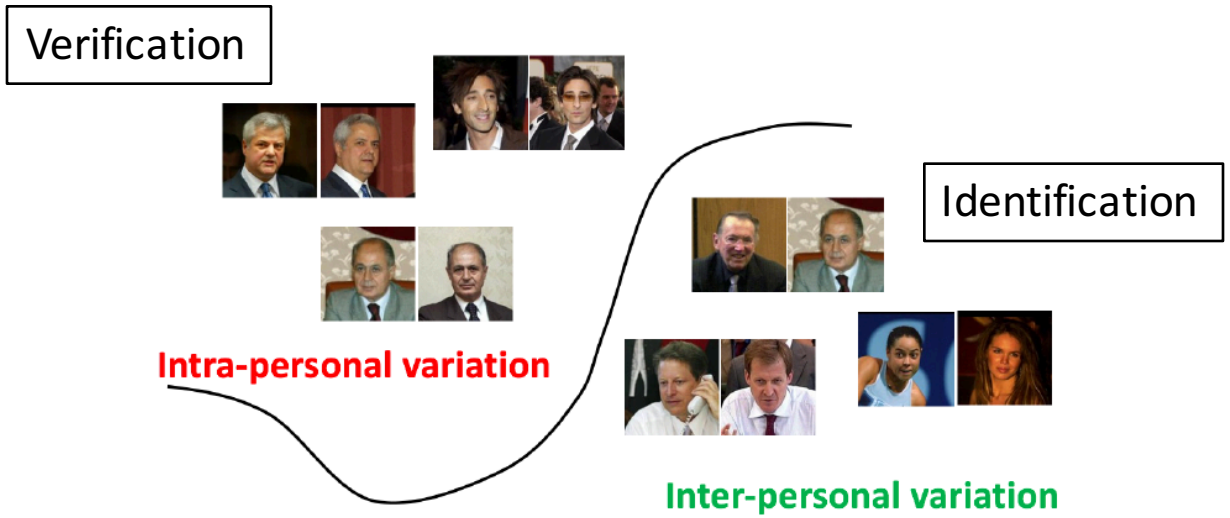
Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$





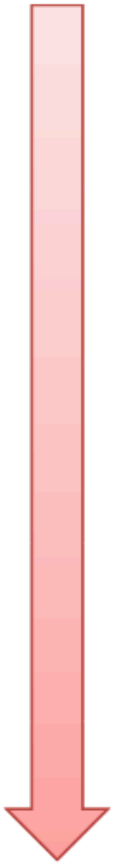
Verification and Identification Signals



- **Identification:**
  - Distinguish images of one identity from another identity
  - Favors large distance between clusters
  - Stronger learning signal, but need not generalize to new identities
- **Verification:**
  - Match two images of an individual across large appearance variations
  - Favors tight clusters for each identity
  - Weaker learning signal, but feature applicable to new identities

Verification and Identification Signals

Learn face representations from



Predicting binary labels (verification)

Prediction becomes richer

Prediction becomes more challenging

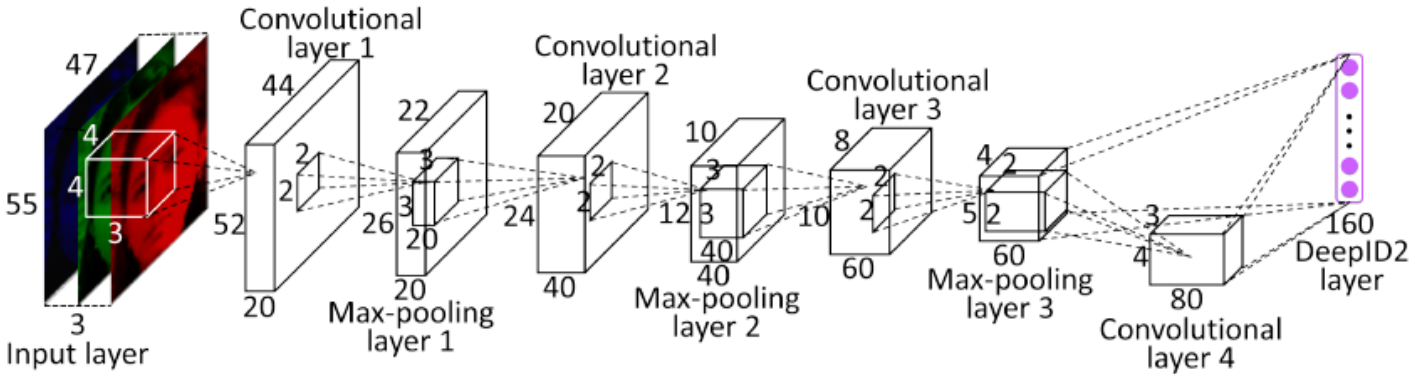
Predicting multi-class labels (identification)

Supervision becomes stronger

Feature learning becomes more effective

Predicting thousands of real-valued pixels (multi-view) reconstruction

Identification Signal



- **Identification: connect feature layer to n-way softmax layer**
  - Outputs a probability distribution over n classes
  - Train with a cross-entropy loss

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

Feature — Target class

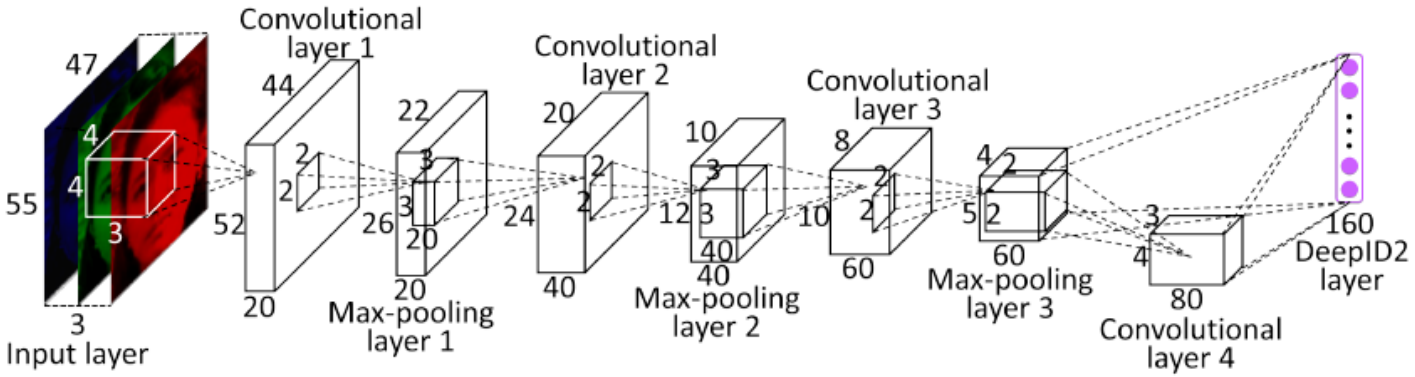
Target probability distribution
p_i = 1 for class t and 0 for other i

Predicted probability distribution

- Goal is to correctly classify all identities simultaneously
  - Incentivize learning discriminative features across inter-personal variations

[Sun et al., DeepID2, NeurIPS 2014]

Verification Signal



- **Verification: directly regularize the feature vector**
  - Pairwise: Gather faces from same class, push those from different classes

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

  - Cosine similarity:

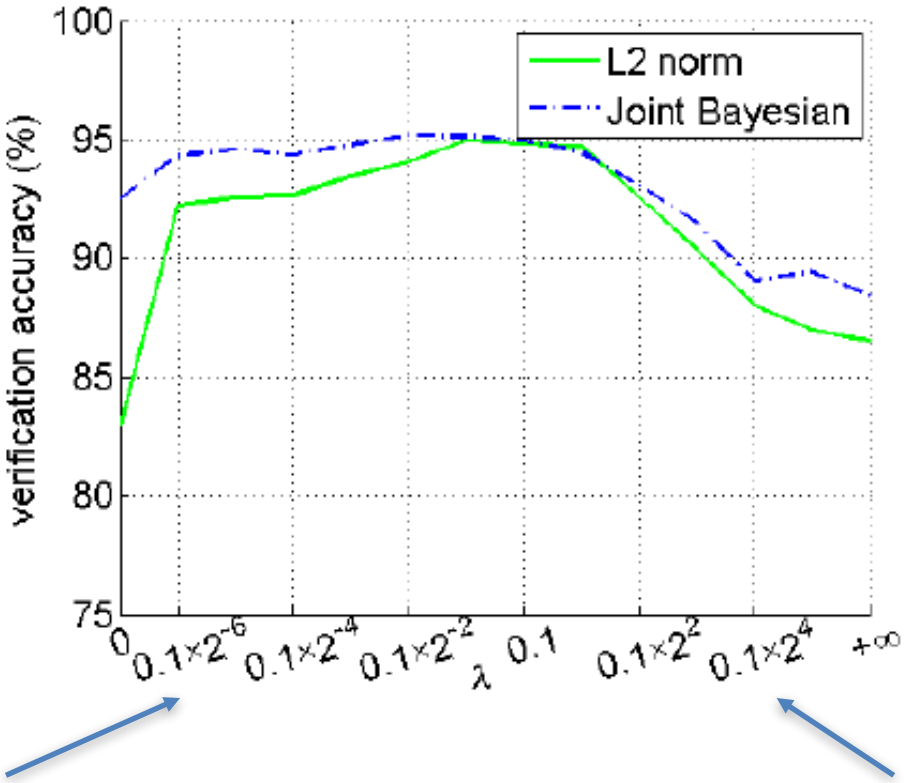
$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \frac{1}{2}\left(y_{ij} - \sigma(wd + b)\right)^2 \text{ , binary } y_{ij}, \quad d = \frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2}$$

- Goal is to learn features that can be matched across intra-personal variations

 [Sun et al., DeepID2, NeurIPS 2014]

Balancing Identification and Verification

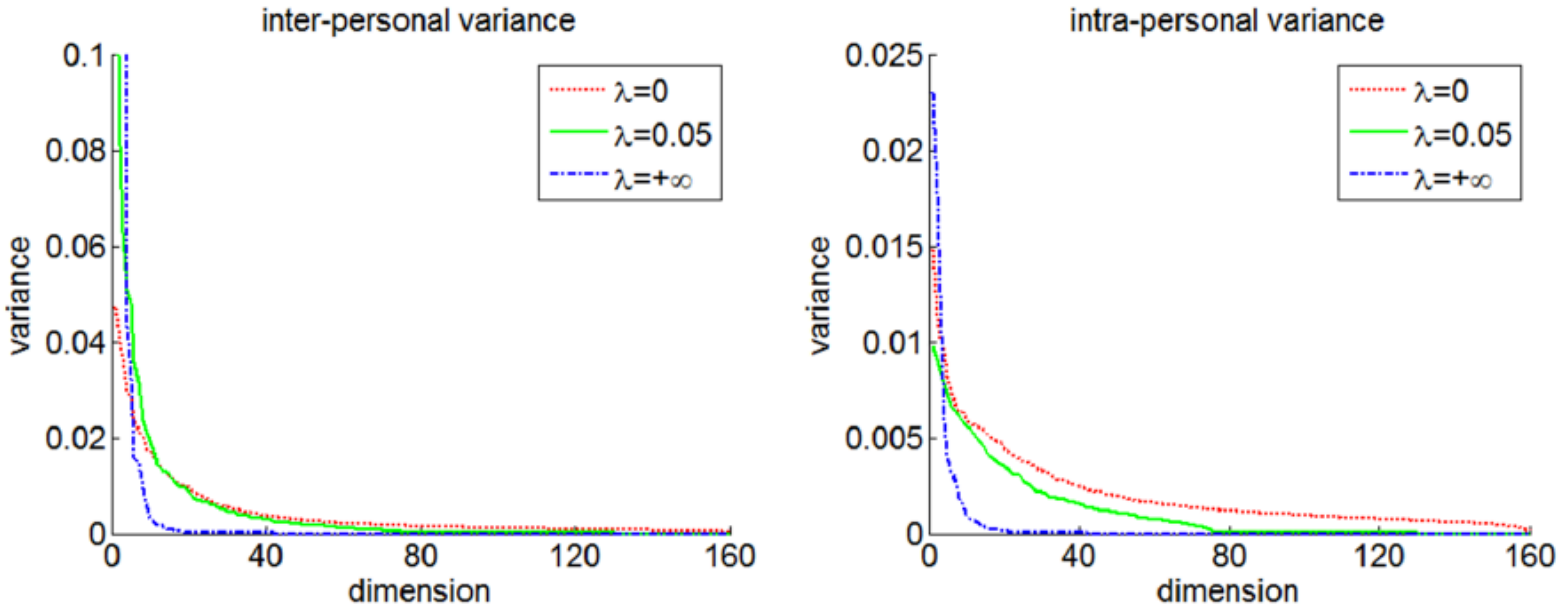
- Balance required between signals to learn good features
- High λ : more weight on verification



Small λ : only identification

Large λ : only verification

Balancing Identification and Verification



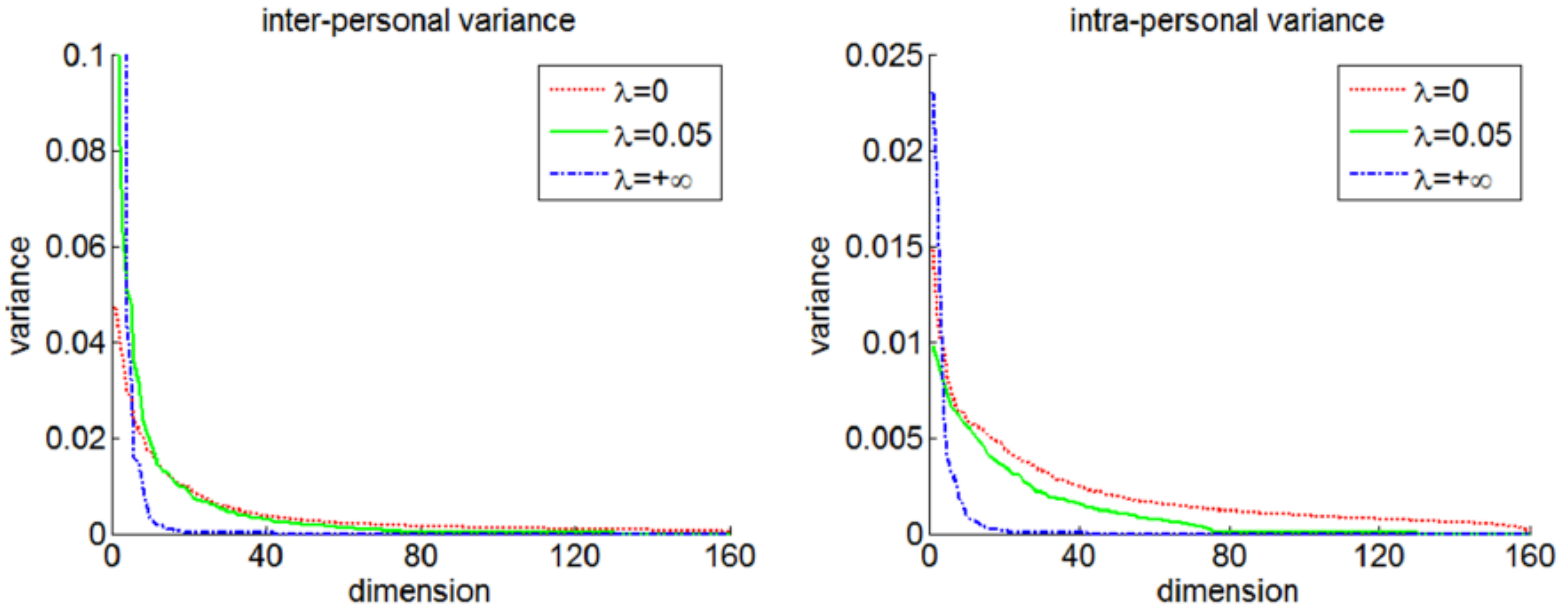
- Inter-class scatter : $\sum_{i=1}^{c} n_i \cdot (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^\top$

c classes

- Intra-class scatter : $\sum_{i=1}^{c} \sum_{x \in D_i} (x - \bar{x}_i)(x - \bar{x}_i)^\top$

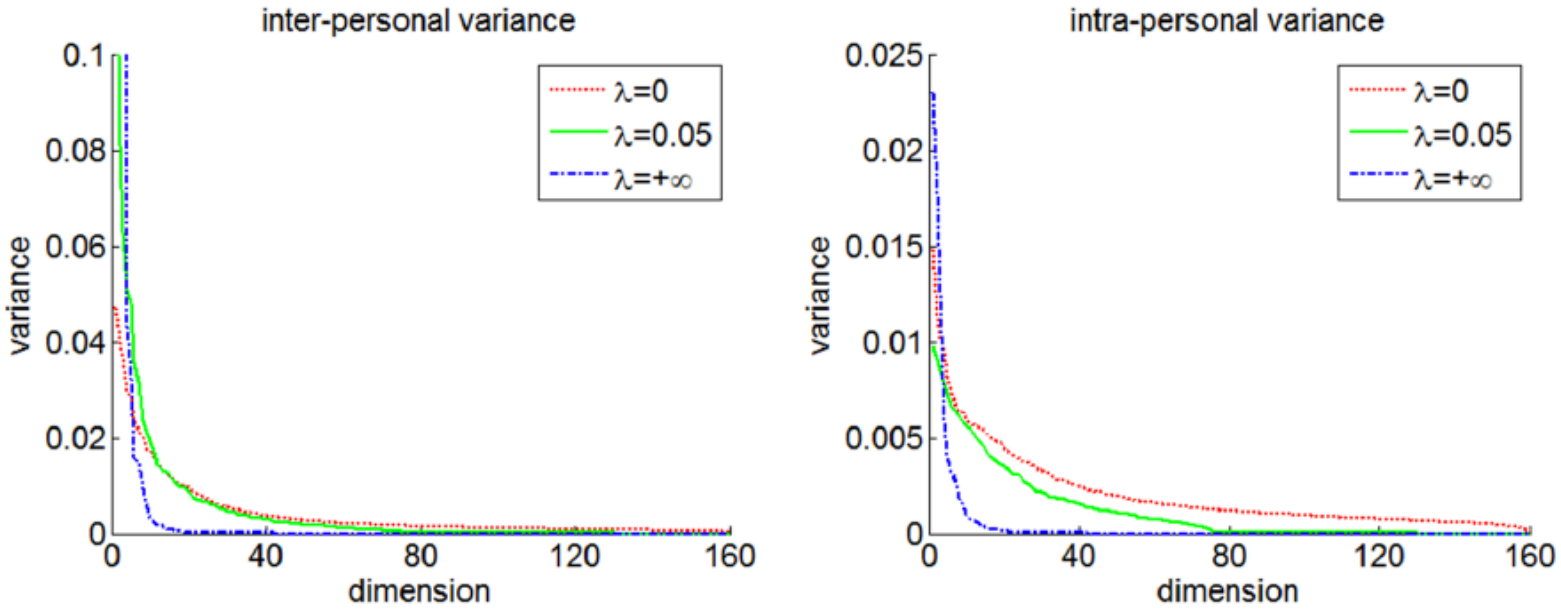
- Variance in scatter indicated by size of eigenvalues
- Small number of eigenvectors: diversity of variation is low
- Both diversity and magnitude of feature variance matters for recognition

Balancing Identification and Verification



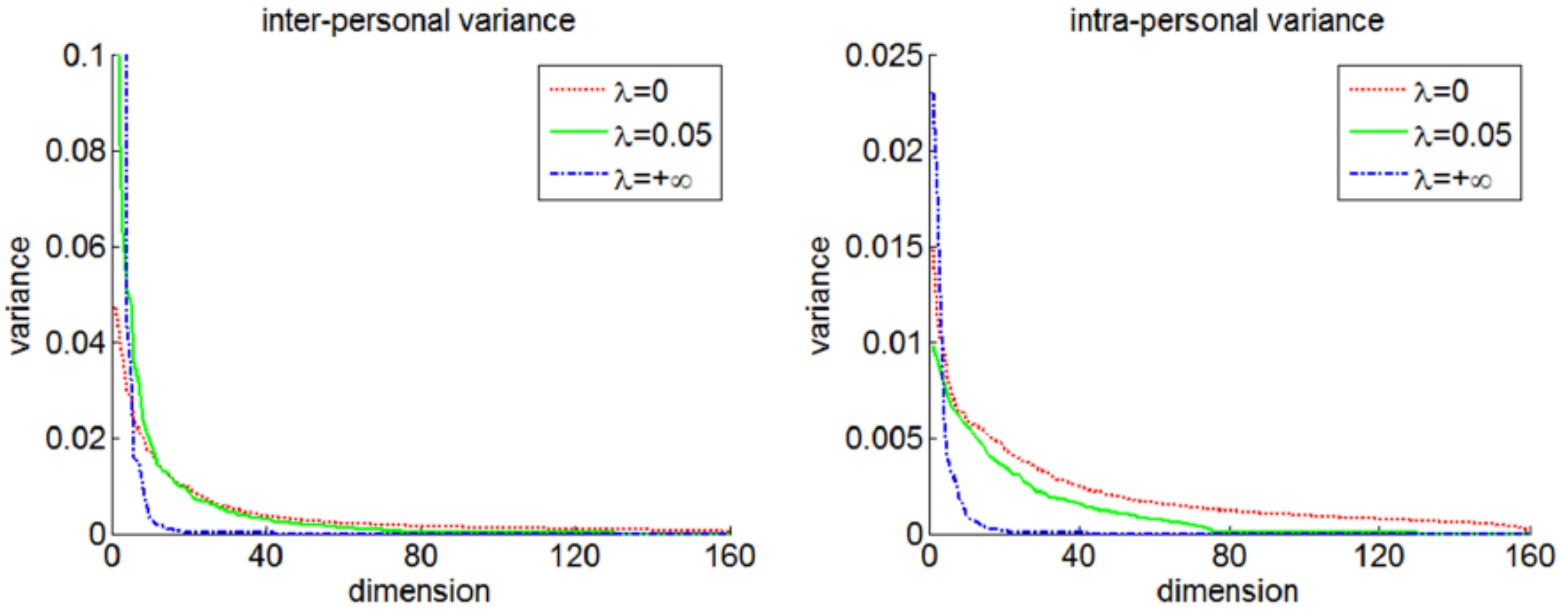
- When only identification signal is used $(\lambda = 0)$:
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification

Balancing Identification and Verification



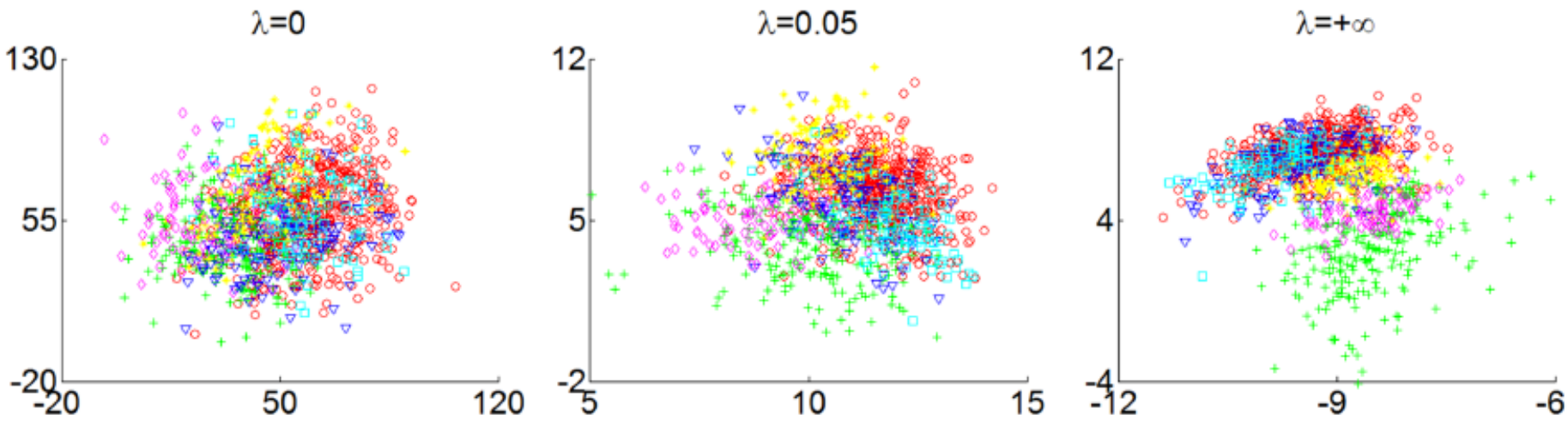
- When only identification signal is used $(\lambda = 0)$:
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification
- When only verification signal is used (λ approaches $+\infty$):
  - Both intra-personal and inter-personal variance collapse to few directions
  - Cannot distinguish between different identities

Balancing Identification and Verification



- When both verification and identification signals are used $(\lambda = 0.05)$:
  - Inter-personal variations stay high
  - Intra-personal variations reduce in diversity and magnitude

Balancing Identification and Verification



- Visualize features for 6 identities
- With only identification signal:
  - Cluster centers are well-separated, but large cluster size causes overlap
- With only verification signal:
  - Cluster sizes become small, but cluster centers also collapse
- With both signals :
  - Clusters sizes become small and cluster centers are reasonably separated

Learning a Face Embedding

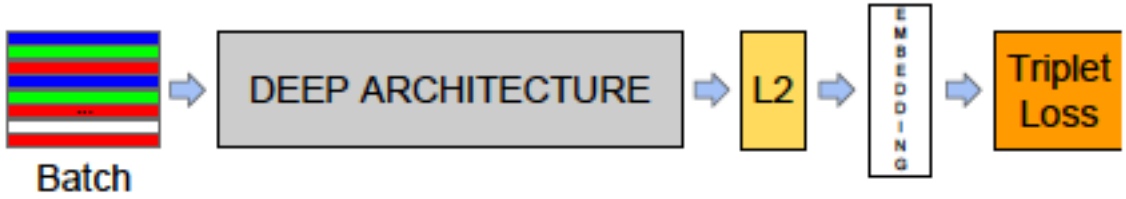
Learn an Embedding for Face Recognition

- **Face verification :** determine whether two images are of the same person
- **Face identification :** determine identity of person in an image
- **Face clustering :** find the same person among a collection of faces

- Train a network such that embedding distances directly represent similarity
  - Faces of same person : small distances
  - Faces of different persons : large distances

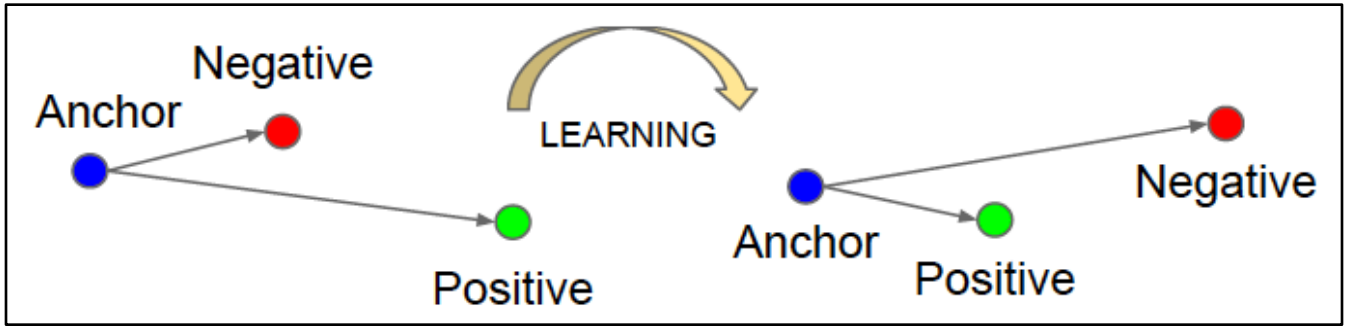
- Once embedding is learned, above problems are all solvable
  - **Verification :** threshold distance between two embeddings
  - **Identification :** can be posed as k-NN classification
  - **Clustering :** can be solved using methods like k-means

Learn an Embedding for Face Recognition



- Goal: learn d-dimensional embedding *f(x)* for face image *x*
- Constrain embedding to lie on unit sphere: $\|f(x)\|_2 = 1$

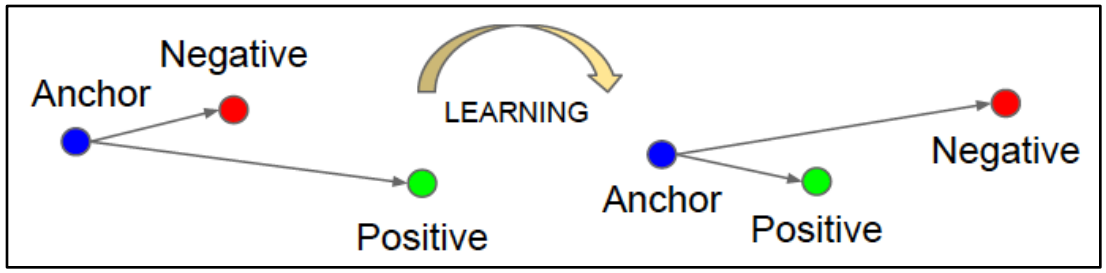
- Goal for triplet loss:
  - Minimize distance between anchor and a positive (from same class)
  - Maximize distance between anchor and a negative (from different classes)



[Schroff et al., FaceNet, CVPR 2015]

Triplet Loss for Training

- Goal for triplet loss:
  - Minimize distance between anchor image x_i^a and a positive x_i^p
  - Maximize distance between anchor x_i^a and a negative x_i^n



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

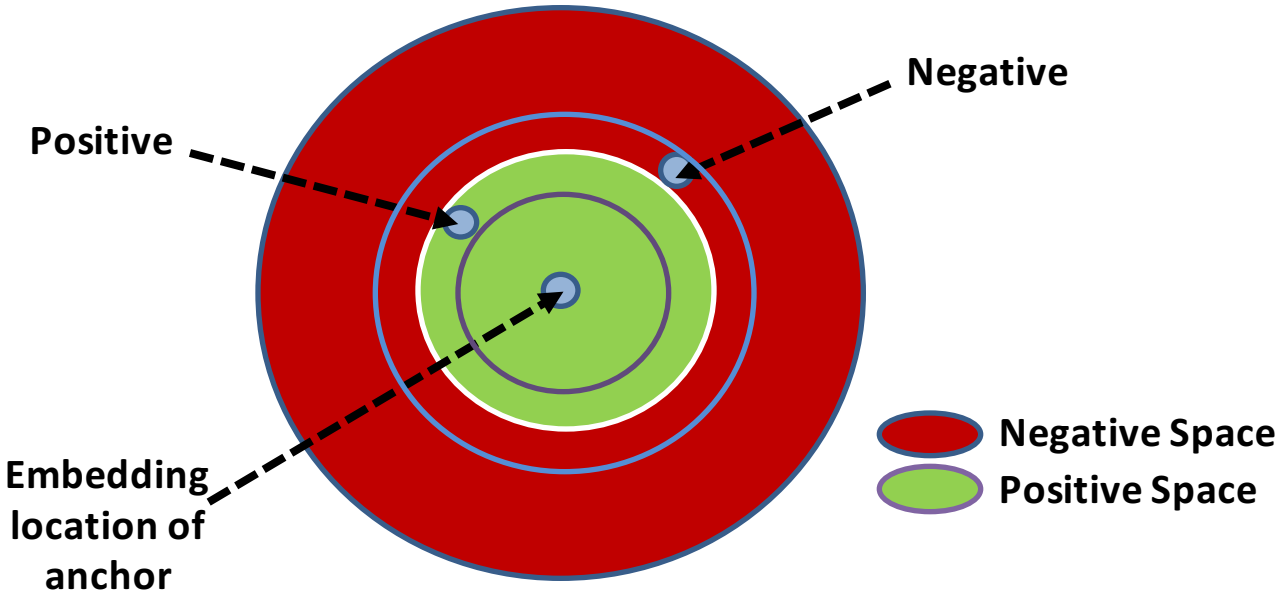
- Total loss to minimize:

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

- Challenge: too many triplets satisfy the margin easily
- Need to select hard examples that are active and improve the model

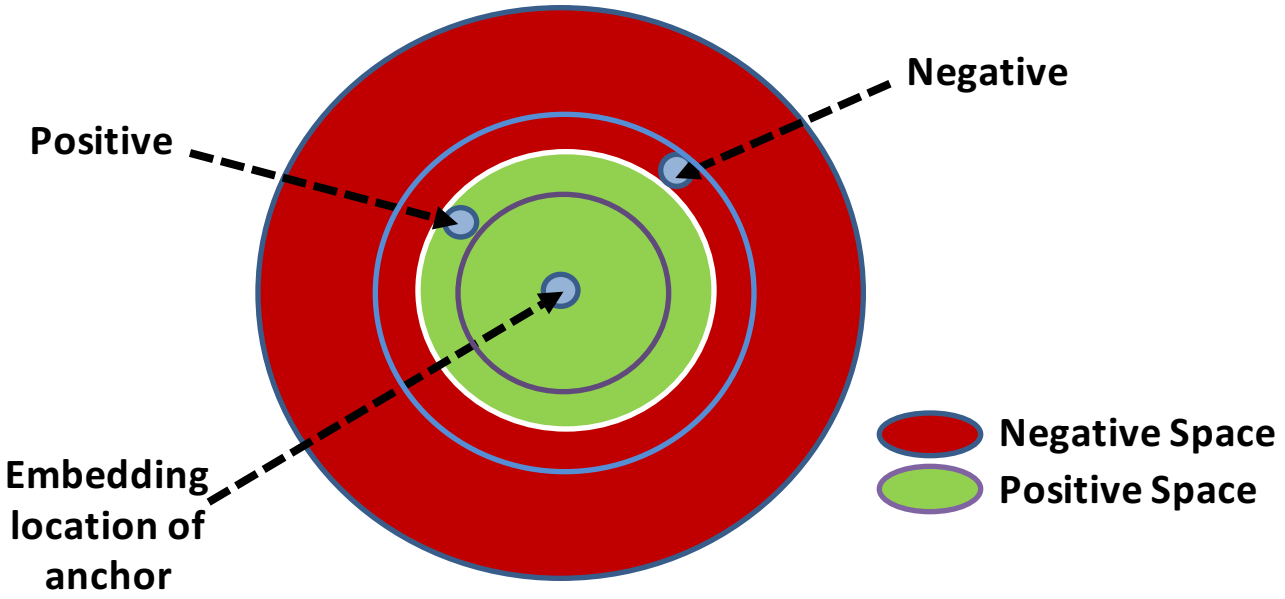
Triplet Selection

- To ensure fast convergence, given an anchor x_i^a :
  - Select **hard positive** x_i^p such that $\text{argmax}_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2$
  - Select **hard negative** x_i^n such that $\text{argmin}_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2$



Triplet Selection

- To ensure fast convergence, given an anchor x_i^a :
  - Select **hard positive** x_i^p such that $\text{argmax}_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2$
  - Select **hard negative** x_i^n such that $\text{argmin}_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2$



- **Inefficient** to compute argmin and argmax over training set
- Might lead to **poor training** as mislabeled or poorly imaged examples dominate

Triplet Selection

- Two courses of action: offline and online selection of triplets
- **Offline:** every n steps, use current feature for argmin and argmax on subset

Triplet Selection

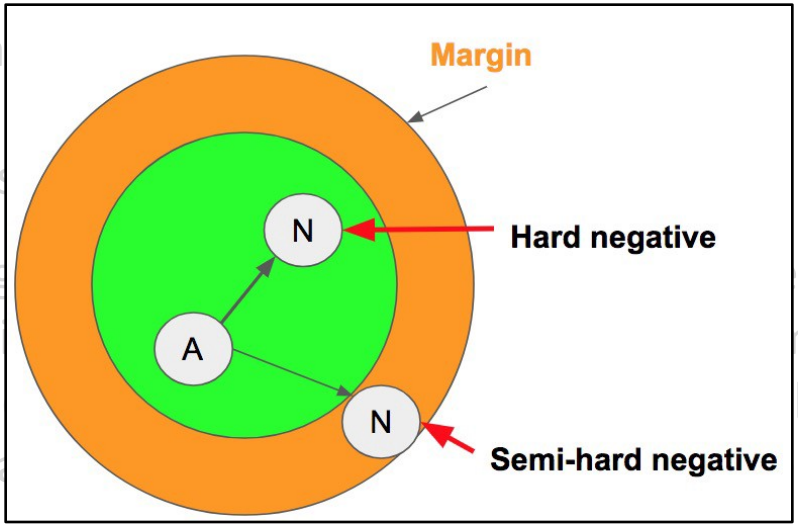
- Two courses of action: offline and online selection of triplets

- **Offline:** every n steps, use current feature for argmin and argmax on subset

- **Online:** selecting hard positive and negative examples in mini-batch
  - Use large mini-batch with several thousand examples
  - Use several examples per identity for meaningful anchor-positive distances
  - Randomly sample negatives from other identities

Triplet Selection

- Two courses of a … triplets
- **Offline:** every n … and argmax on subset
- **Online:** selecting … s in mini-batch
  - Use large m… nples
  - Use several … l anchor-positive distances
  - Randomly s… s

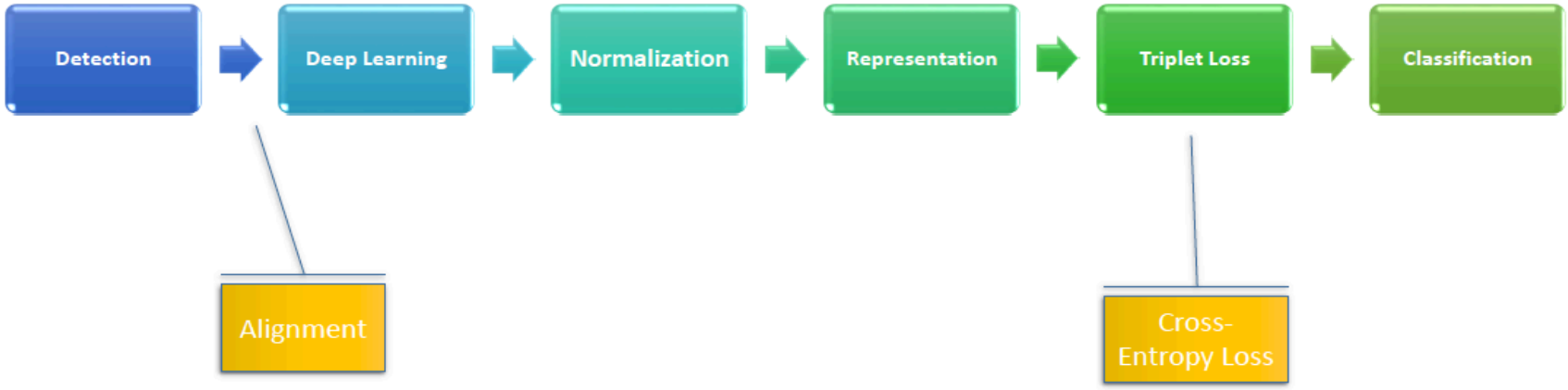


- In practice:
  - Use all anchor-positive pairs, instead of just hard positives
  - Use **semi-hard negatives** at beginning of training

$$\|f(x_i^a) - f(x_i^p)\|_2^2 < \|f(x_i^a) - f(x_i^n)\|_2^2$$

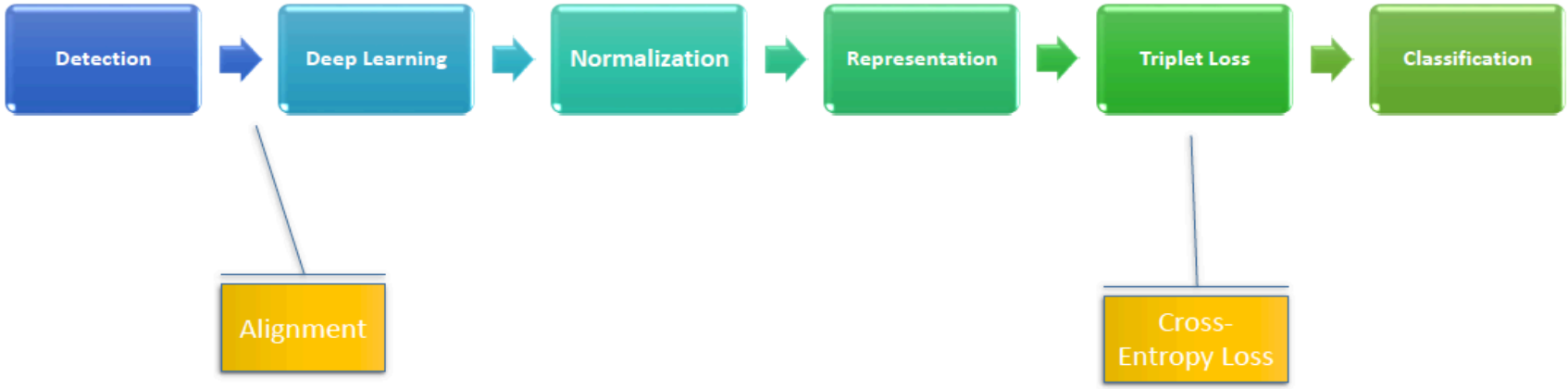
  - Not hardest negatives, can be within margin, but further than positives
  - Hardest negatives at beginning can cause feature to collapse

Comparison of DeepFace, DeepID2, FaceNet



- Benefit over DeepFace:
  - Learns an embedding, can be used for multiple tasks
  - Only 128-dimensional representation, efficient for inference

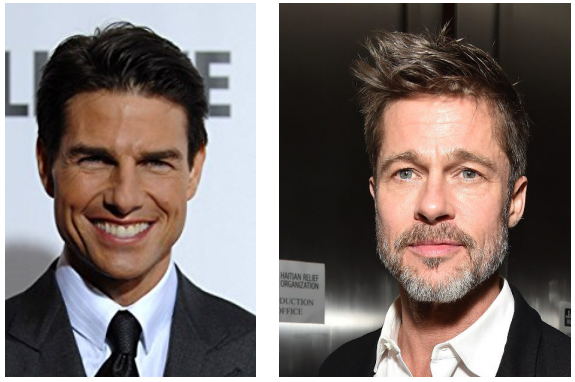
Comparison of DeepFace, DeepID2, FaceNet

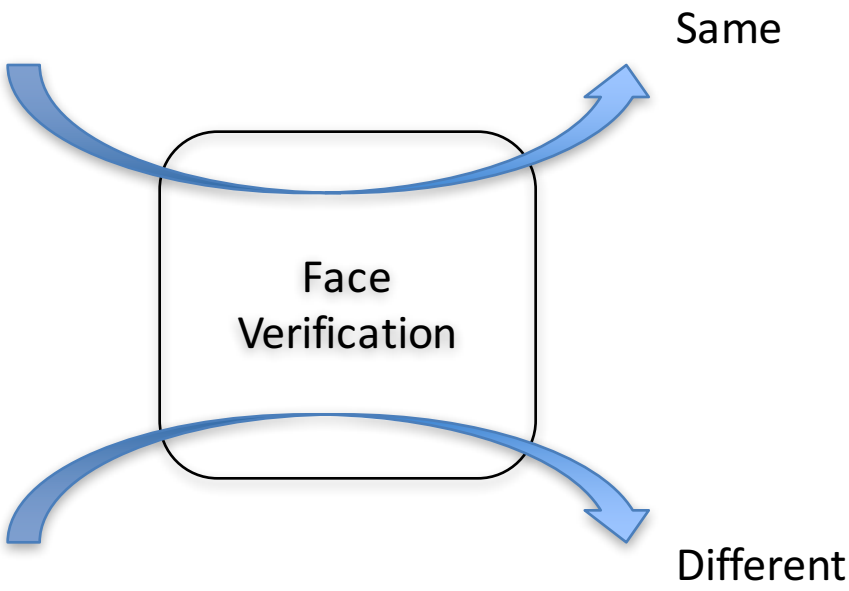


- Benefit over DeepFace:
  - Learns an embedding, can be used for multiple tasks
  - Only 128-dimensional representation, efficient for inference
- Intuitive benefit of triplet loss over pairwise loss in DeepID2
  - Pairwise: map all faces from one identity to same point
  - Triplet: margin between each pair of faces of an identity and all other faces
  - Triplet loss allows an identity manifold, with distance from other identities

Evaluation of Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$





Evaluation of Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$

- True accepts: set of face pairs correctly classified as same within a threshold *d*

$$\mathrm{TA}(d) = \{(i,j) \in \mathcal{P}_{\text{same}}, \text{with } D(x_i, x_j) \leq d\}$$

All face pairs in test set that belong to same identities

Evaluation of Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$

- True accepts: set of face pairs correctly classified as same within a threshold *d*

$$\mathrm{TA}(d) = \{(i,j) \in \mathcal{P}_{\text{same}}, \text{with } D(x_i, x_j) \leq d\}$$

All face pairs in test set that belong to same identities

- False accepts: set of face pairs incorrectly classified as same for threshold *d*

$$\mathrm{FA}(d) = \{(i,j) \in \mathcal{P}_{\text{diff}}, \text{with } D(x_i, x_j) \leq d\}$$

All face pairs in test set that belong to different identities

Evaluation of Face Verification

- Given a pair of face images:
  - A squared L2 distance $D(x_i, x_j)$ is used to determine same or different
  - Good embedding: true matches will lie within a small value of $D(x_i, x_j)$

- True accepts: set of face pairs correctly classified as same within a threshold *d*

$$\mathrm{TA}(d) = \{(i,j) \in \mathcal{P}_{\text{same}}, \text{with } D(x_i, x_j) \leq d\}$$

All face pairs in test set that belong to same identities

- False accepts: set of face pairs incorrectly classified as same for threshold *d*

$$\mathrm{FA}(d) = \{(i,j) \in \mathcal{P}_{\text{diff}}, \text{with } D(x_i, x_j) \leq d\}$$

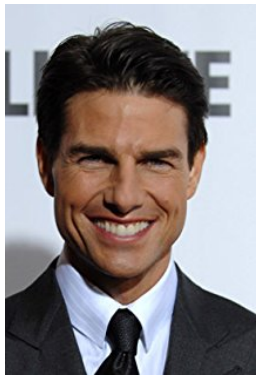
All face pairs in test set that belong to different identities

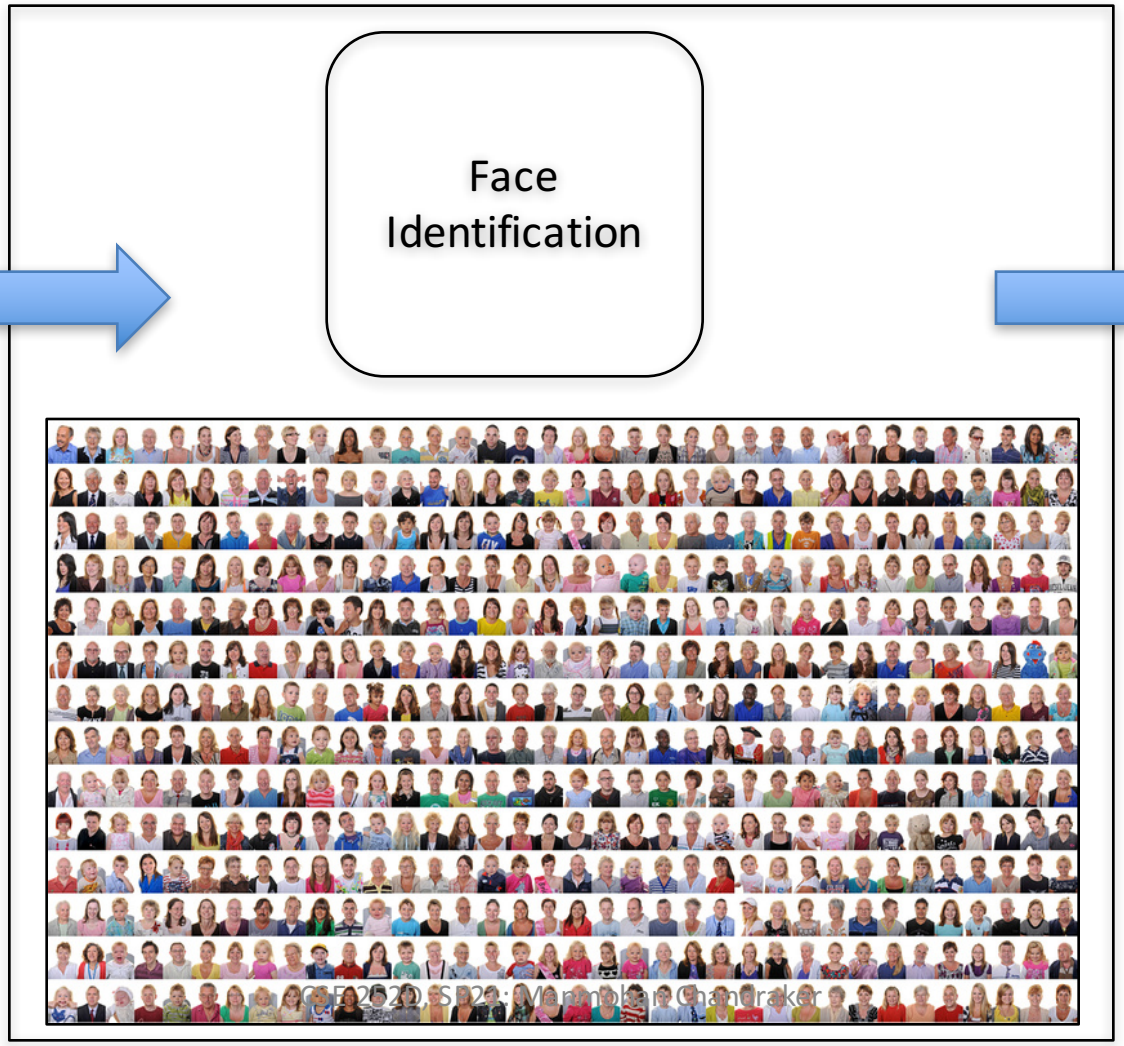
- Determine validation rate and false accept rate:

$$\mathrm{VAL}(d) = \frac{|\mathrm{TA}(d)|}{|\mathcal{P}_{\text{same}}|}, \quad \mathrm{FAR}(d) = \frac{|\mathrm{FA}(d)|}{|\mathcal{P}_{\text{diff}}|}$$

Identification Harder than Verification

- Closed set identification: assign probe image one of gallery identities
- Galleries can be very large, high chance of similar appearances





Identification Harder than Verification

- Closed set identification: assign probe image one of gallery identities
- Galleries can be very large, high chance of similar appearances

- True accept rate: percentage of probes matched correctly to gallery
- False accept rate: percentage of probes matched incorrectly to gallery
- Aim: achieve high TAR at low FAR

Identification Harder than Verification

- Closed set identification: assign probe image one of gallery identities
- Galleries can be very large, high chance of similar appearances

- True accept rate: percentage of probes matched correctly to gallery
- False accept rate: percentage of probes matched incorrectly to gallery
- Aim: achieve high TAR at low FAR

- Suppose a verification system achieves 99% accuracy
- Incorrectly verifies 2 cases among 100 matched and 100 unmatched pairs
- Now use for identification against gallery of 901 subjects
- Expect 1 correct and 9 incorrect candidates, all of which look similar
- Even if resolve further by 50%, still have a 50% error rate

Identification Harder than Verification

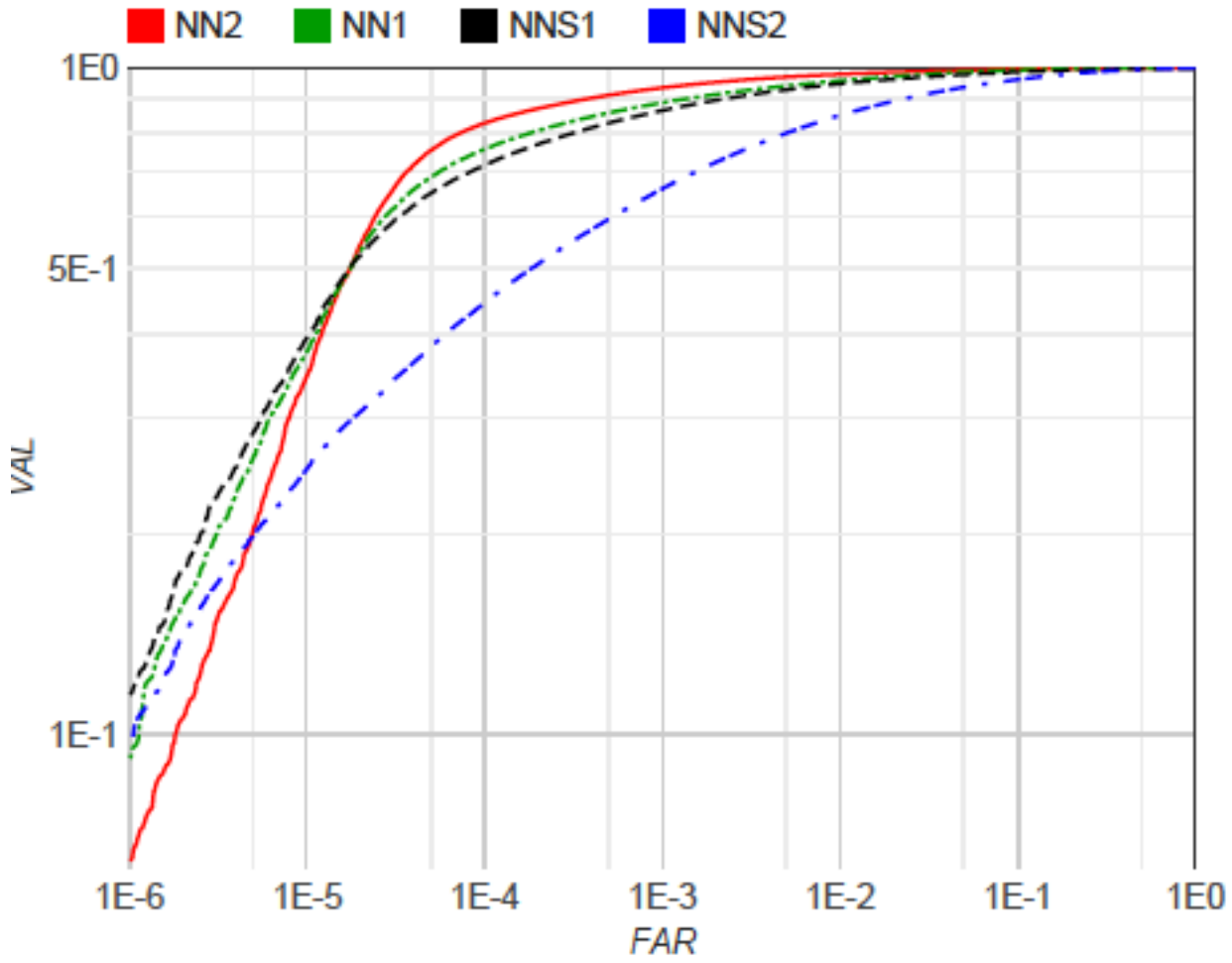
- Closed set identification: assign probe image one of gallery identities
- Galleries can be very large, high chance of similar appearances

- True accept rate: percentage of probes matched correctly to gallery
- False accept rate: percentage of probes matched incorrectly to gallery
- Aim: achieve high TAR at low FAR

- Suppose a verification system achieves 99% accuracy
- Incorrectly verifies 2 cases among 100 matched and 100 unmatched pairs
- Now use for identification against gallery of 901 subjects
- Expect 1 correct and 9 incorrect candidates, all of which look similar
- Even if resolve further by 50%, still have a 50% error rate

- Even harder case: open set identification
  - Probe identity may or may not exist in the gallery

Evaluation of Face Verification



For a few different network architectures

[Schroff et al., FaceNet, CVPR 2015]

Evaluation of Face Verification

Validation rates at threshold 0.001

jpeg q	val-rate
10	67.3%
20	81.4%
30	83.9%
50	85.5%
70	86.1%
90	86.5%

Image quality

#pixels	val-rate
1,600	37.8%
6,400	79.5%
14,400	84.5%
25,600	85.7%
65,536	86.4%

Image resolution

#dims	VAL
64	$86.8\% \pm 1.7$
128	$87.9\% \pm 1.9$
256	$87.7\% \pm 1.9$
512	$85.6\% \pm 2.0$

Embedding dimension

#training images	VAL
2,600,000	76.3%
26,000,000	85.1%
52,000,000	85.1%
260,000,000	86.2%

Training images

[Schroff et al., FaceNet, CVPR 2015]

Qualitative Results