CSE 252D: Advanced Computer Vision

Manmohan Chandraker

Lecture 9: Face Recognition

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 Kaltura
  - 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 Fri, Apr 30

- Deep Fundamental Matrix Estimation
  - https://openaccess.thecvf.com/content_ECCV_2018/html/Rene_Ranftl_Deep_Fundamental_Matrix_ECCV_2018_paper.html
- DSAC - Differentiable RANSAC for Camera Localization
  - https://arxiv.org/abs/1611.05705
- Unsupervised Monocular Depth Estimation with Left-Right Consistency
  - https://arxiv.org/abs/1609.03677
- LSD-SLAM: Large-Scale Direct Monocular SLAM
  - https://vision.in.tum.de/_media/spezial/bib/engel14eccv.pdf

Papers for Wed, May 05

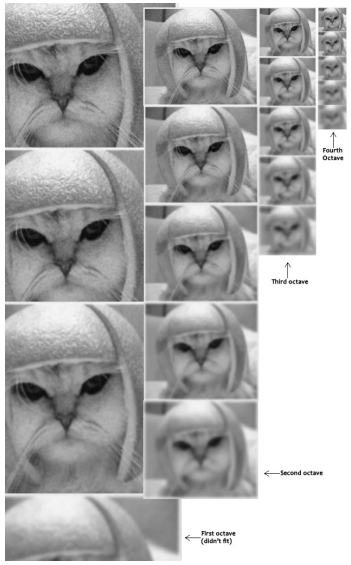
- A Discriminative Feature Learning Approach for Deep Face Recognition
  - https://ydwen.github.io/papers/WenECCV16.pdf
- SphereFace: Deep Hypersphere Embedding for Face Recognition
  - https://arxiv.org/abs/1704.08063
- ArcFace: Additive Angular Margin Loss for Deep Face Recognition
  - https://arxiv.org/abs/1801.07698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

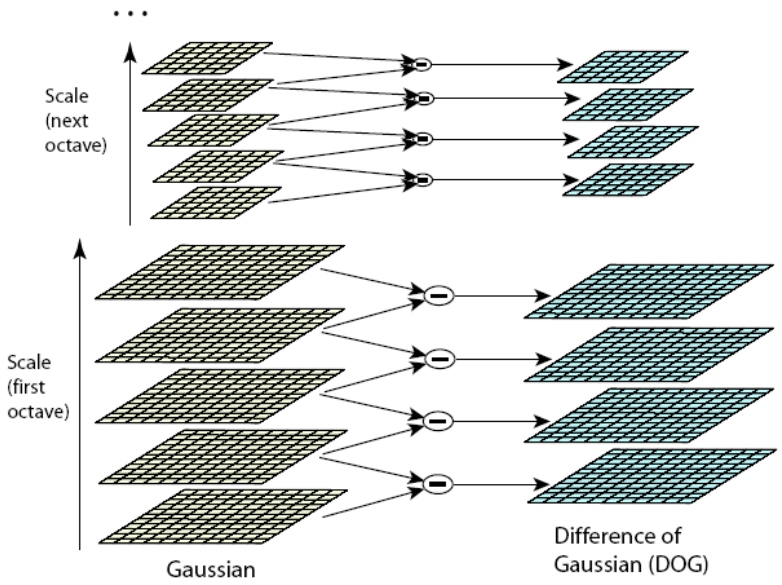
Papers for Fri, May 07

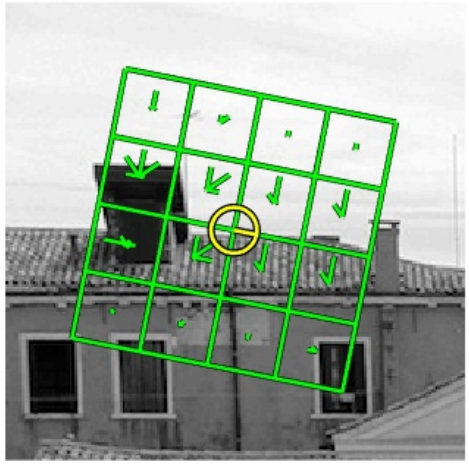
- Deep High-Resolution Representation Learning for Human Pose Estimation
  - https://arxiv.org/abs/1902.09212
- Simple Baselines for Human Pose Estimation and Tracking
  - https://arxiv.org/abs/1804.06208
- OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields
  - https://arxiv.org/abs/1812.08008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

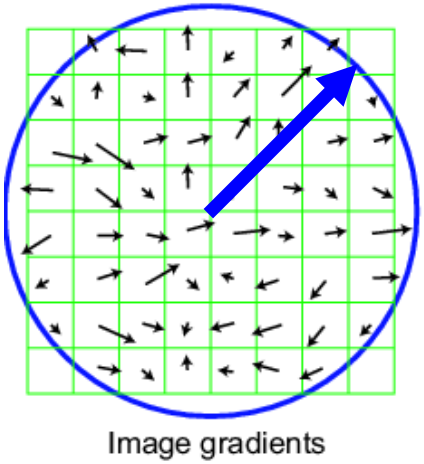
Recap

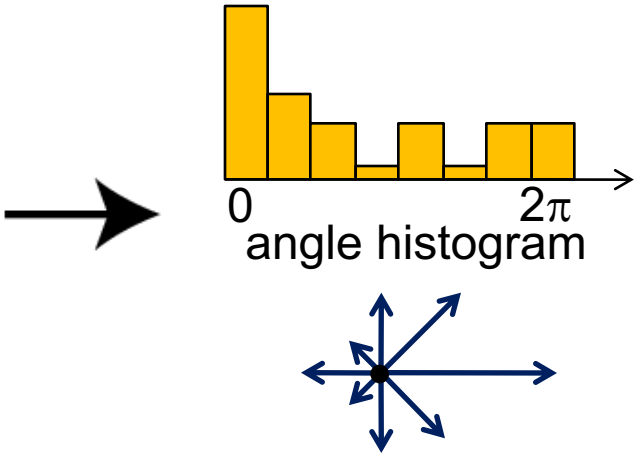
Keypoints and Correspondence

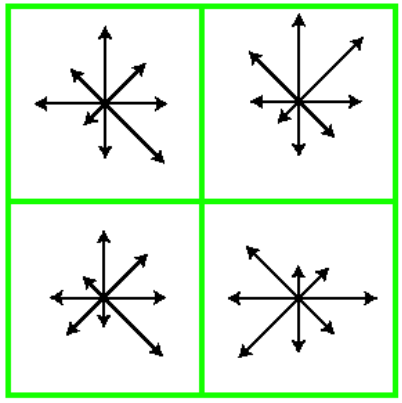




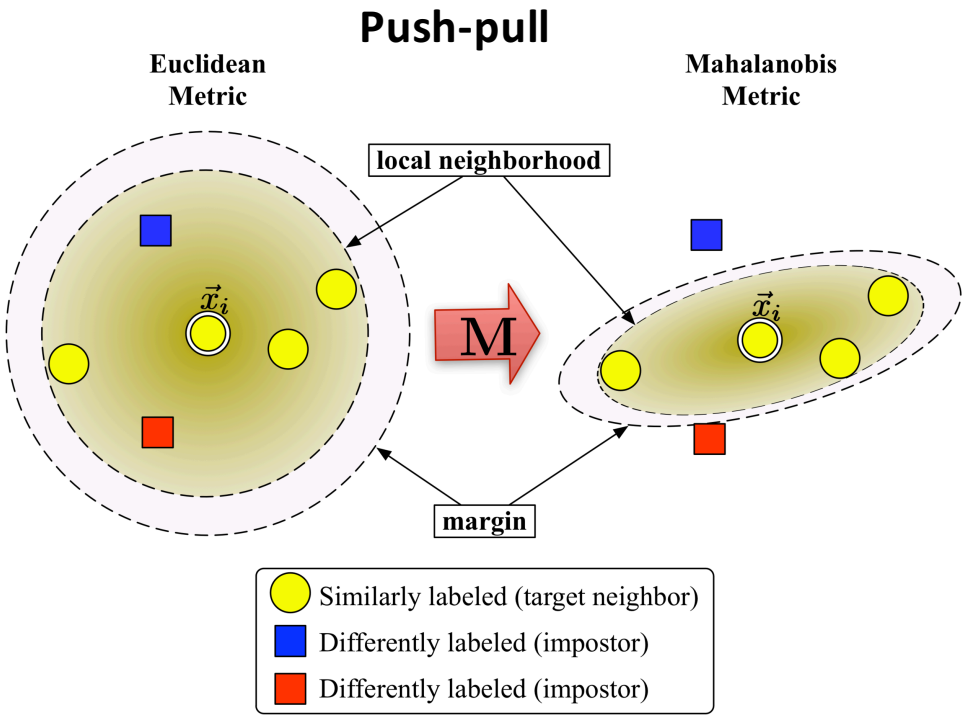


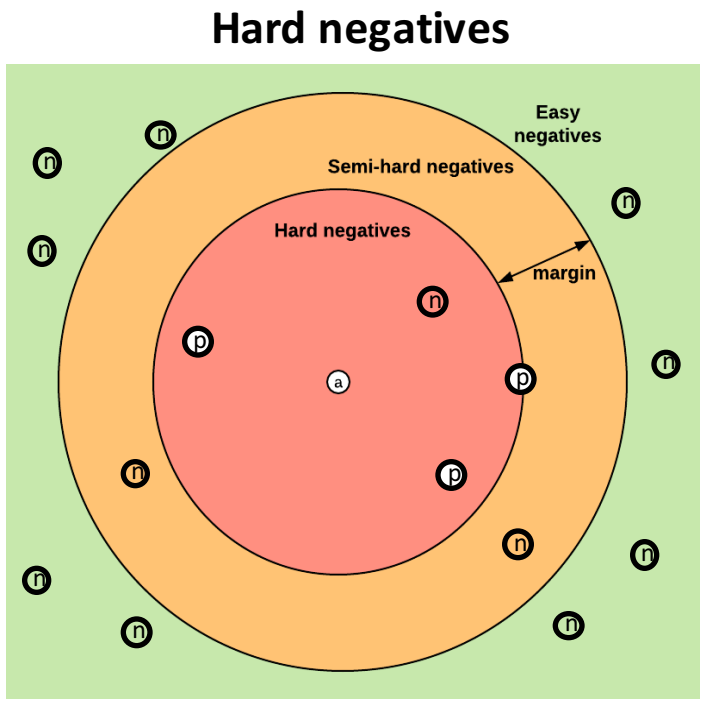






Metric Learning





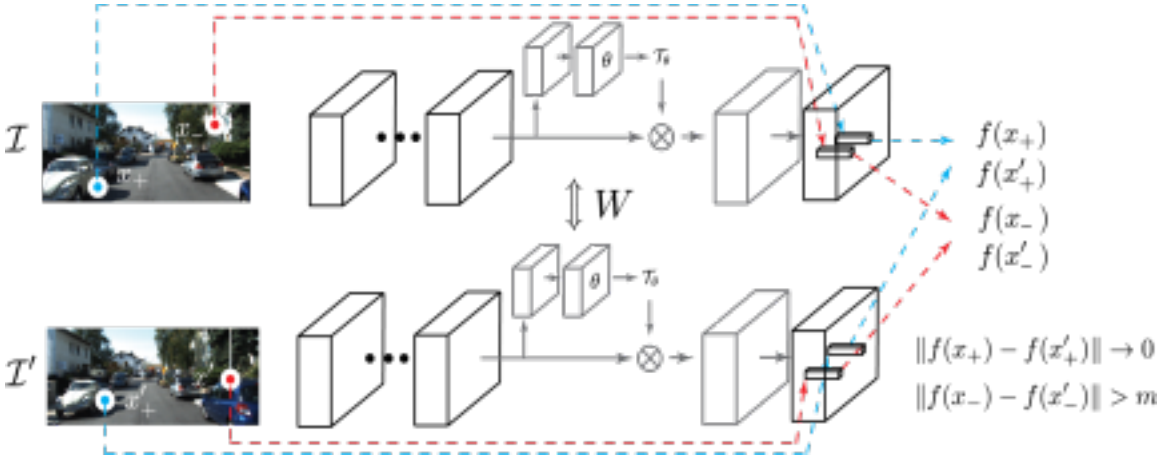
- Contrastive loss:

$$L(x_1, x_2) = s_{12}\ |x_1 - x_2|^2 + (1 - s_{12})\ \max(0, m^2\ - |x_1 - x_2|^2)$$

- Triplet loss:

$$l_{tri}(\mathcal{T}) = \left[\|\mathbf{x}_a - \mathbf{x}_p\|^2 - \|\mathbf{x}_a - \mathbf{x}_n\|^2 + m\right]_+$$

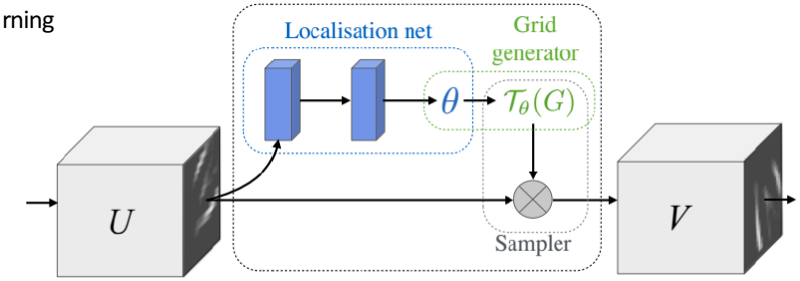
Deep Keypoints and Correspondence



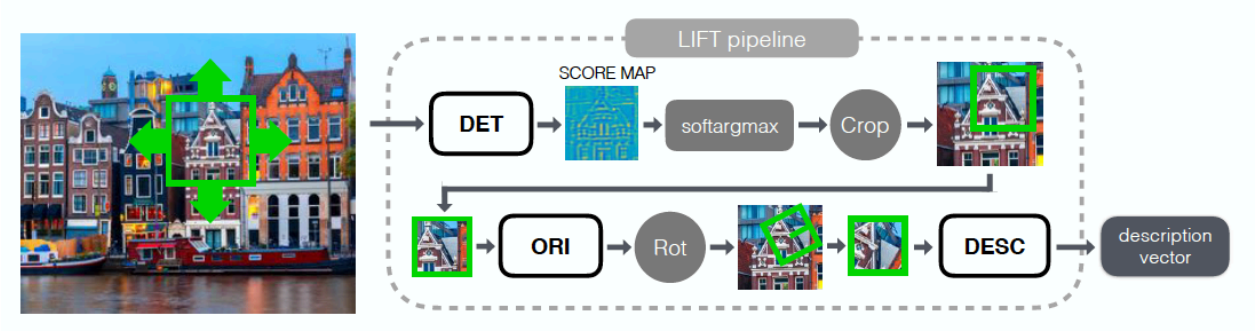
Dense feature extraction

Spatial normalization

Metric learning



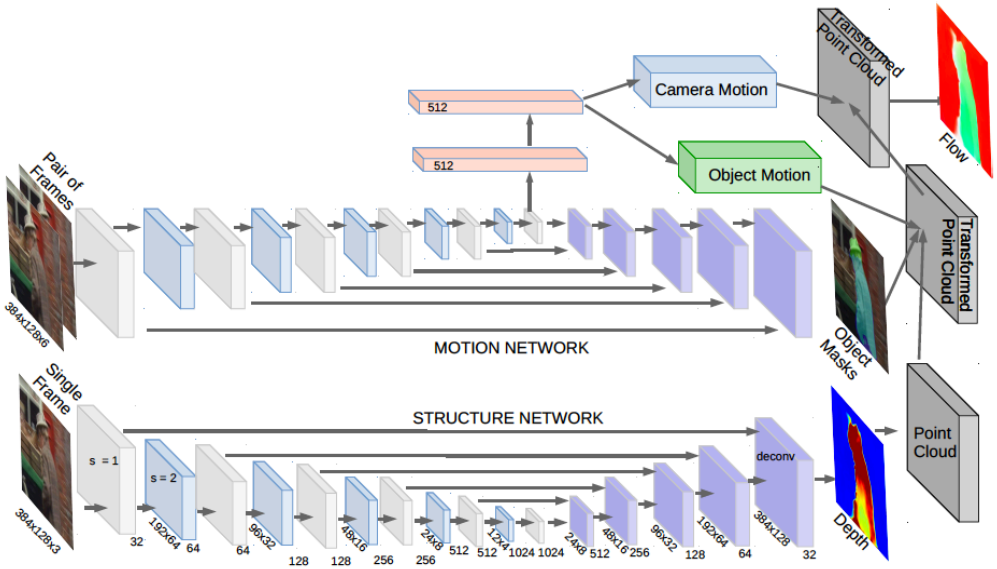
Spatial Transformer



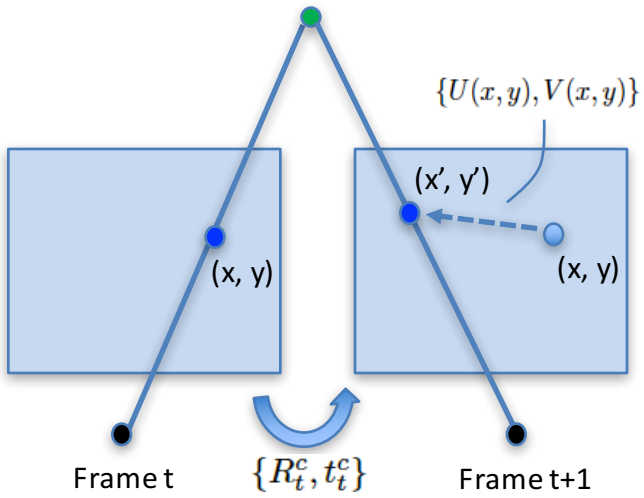
Deep Networks for SFM

- Estimate depths (convert to 3D points given calibration) in frame t
- Estimate motion from frame t to t+1 for background and objects
- Project estimated 3D points to frame t+1 using the estimated motions
- Use a consistency condition to declare matches as good

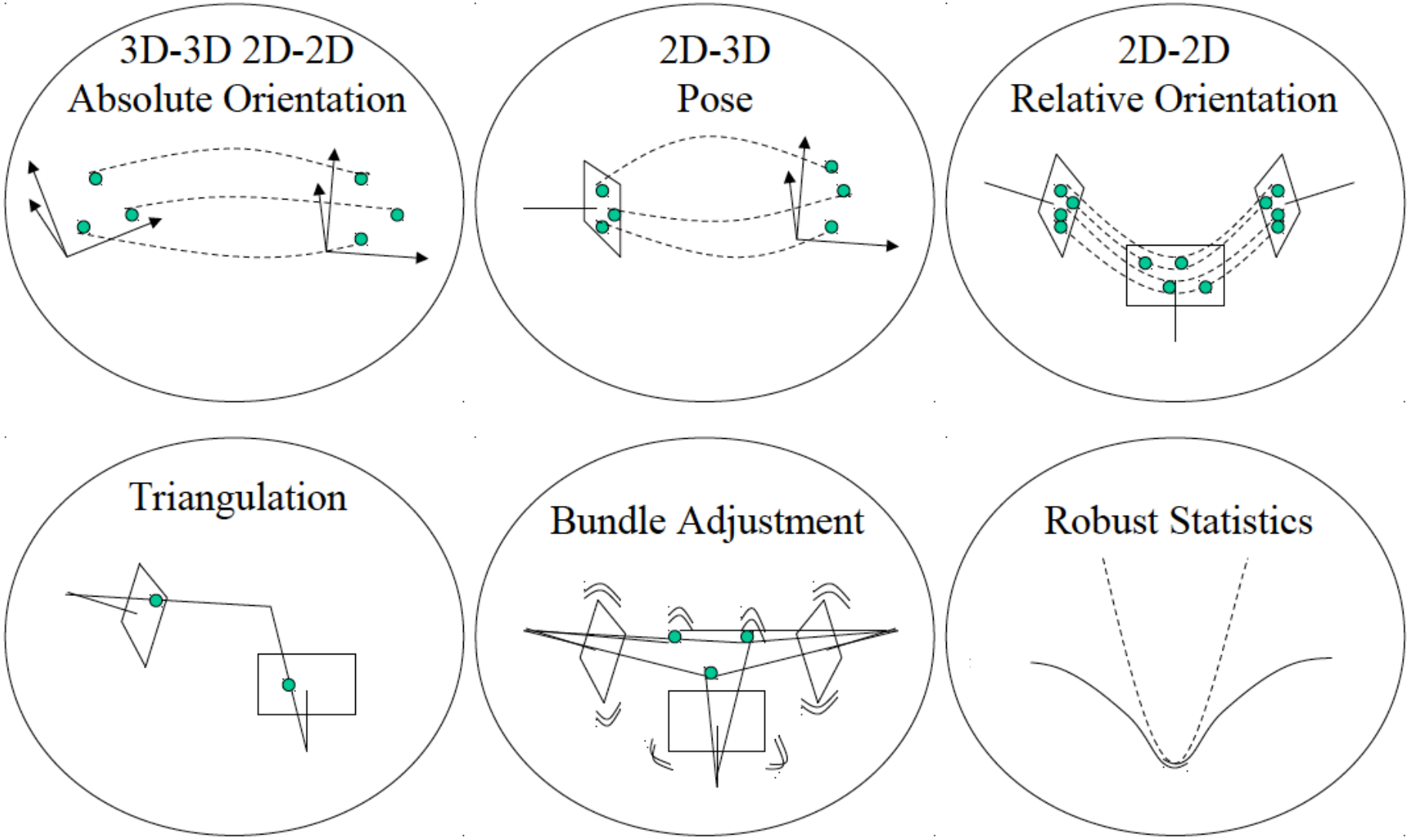
Structure and motion sub-networks



Photoconsistency

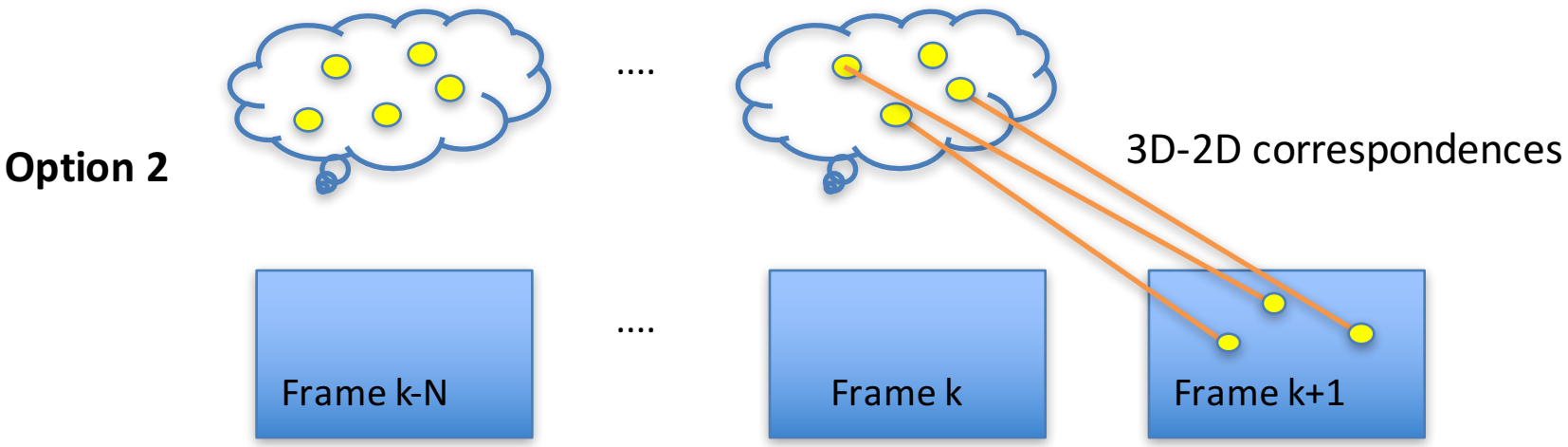


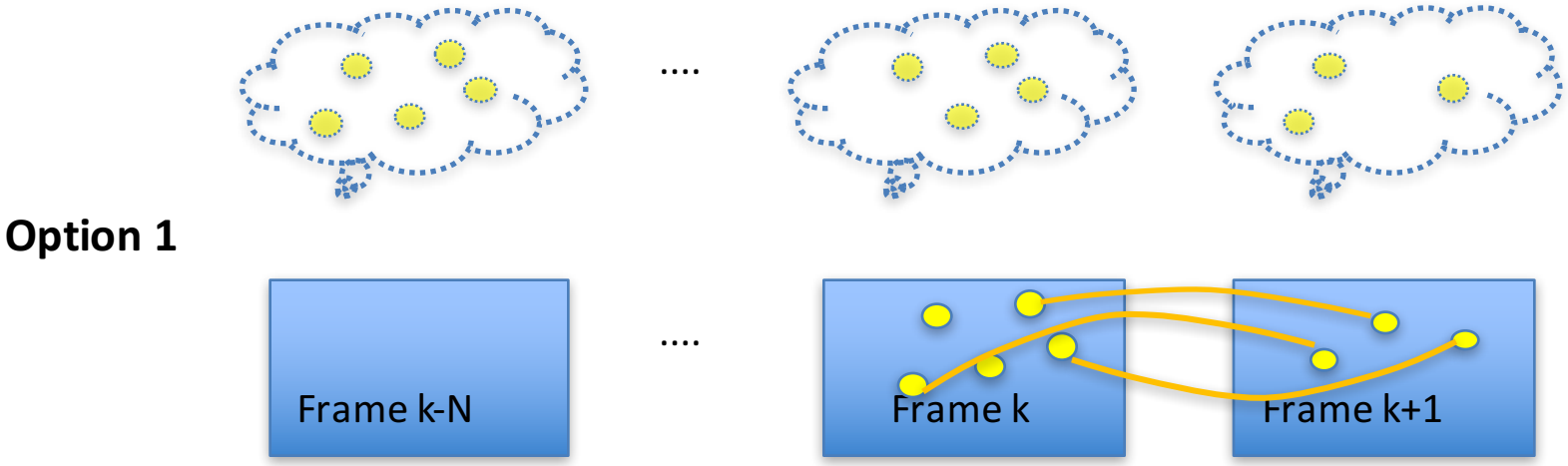
Toolkit for Practical SFM



Real-Time SFM: Two Options

- Usually absolute pose estimations rather than relative pose

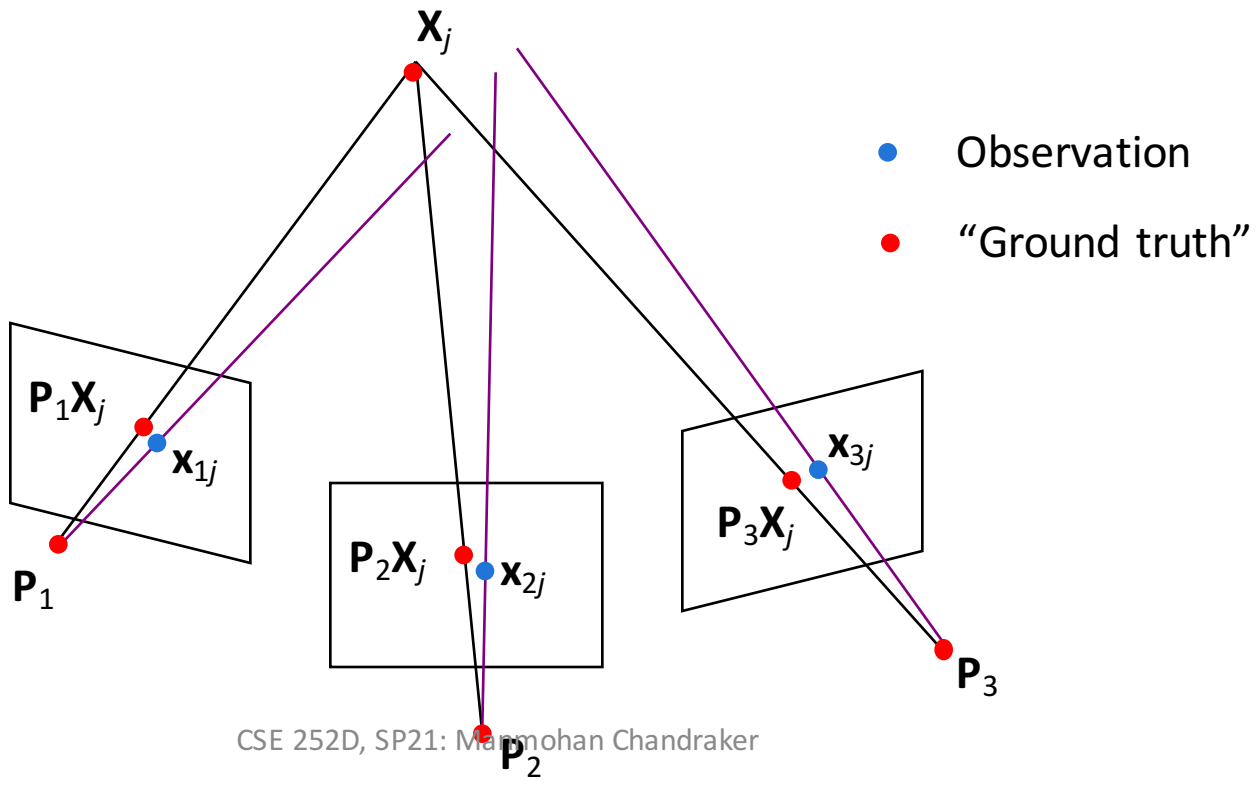




Bundle Adjustment

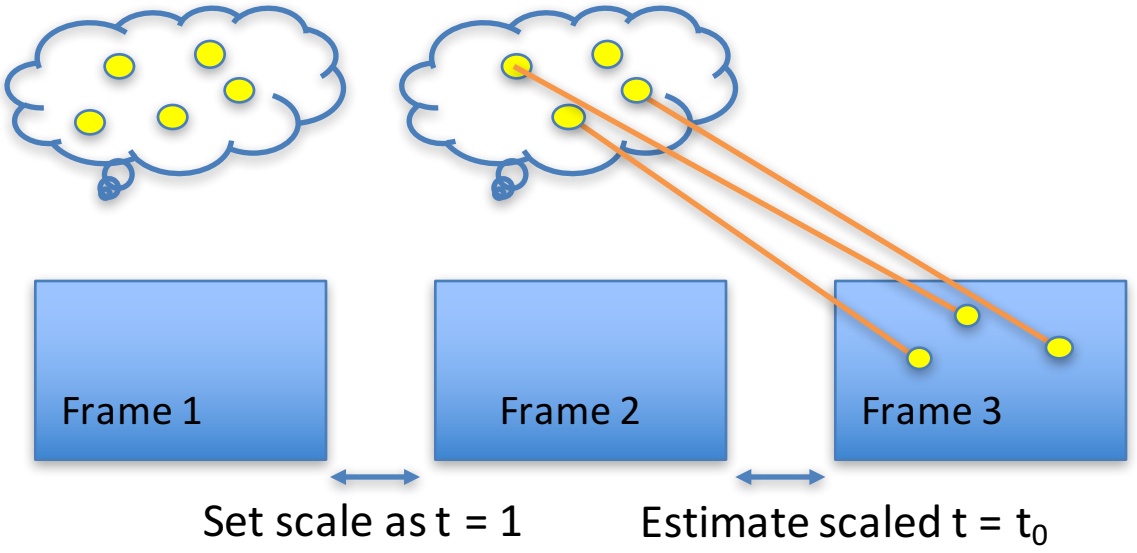
- Non-linear method for refining structure and motion
- Minimize reprojection error

$$Error\,(P,X) = \sum_{Cameras\ P_i} \sum_{Points\ X_j} |x_{ij} - P_i X_j|^2$$

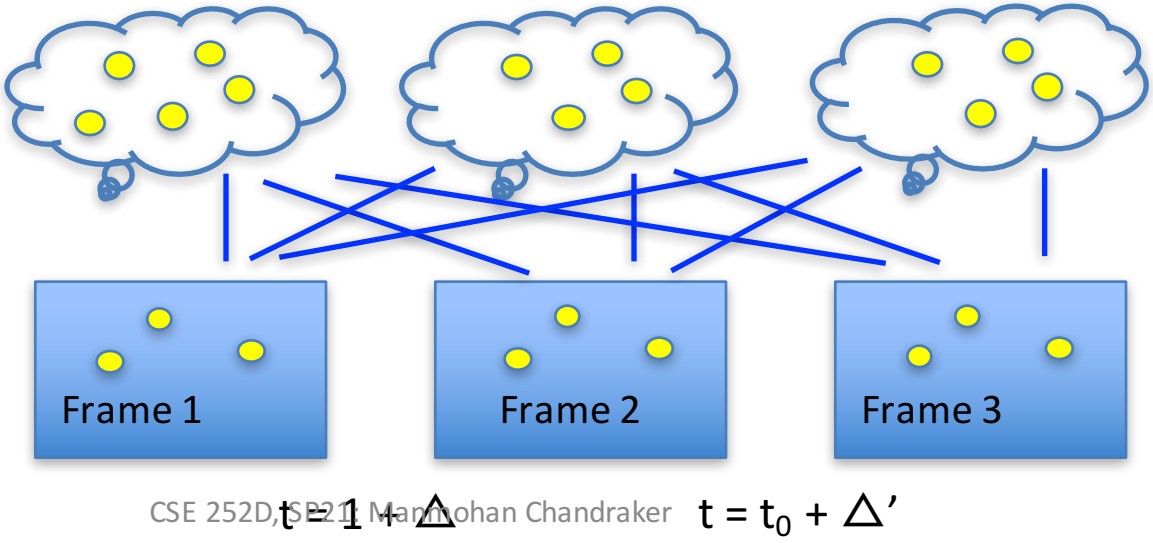


Scale Drift Correction

Choose scale arbitrarily, for example, translation from frame 1 to 2

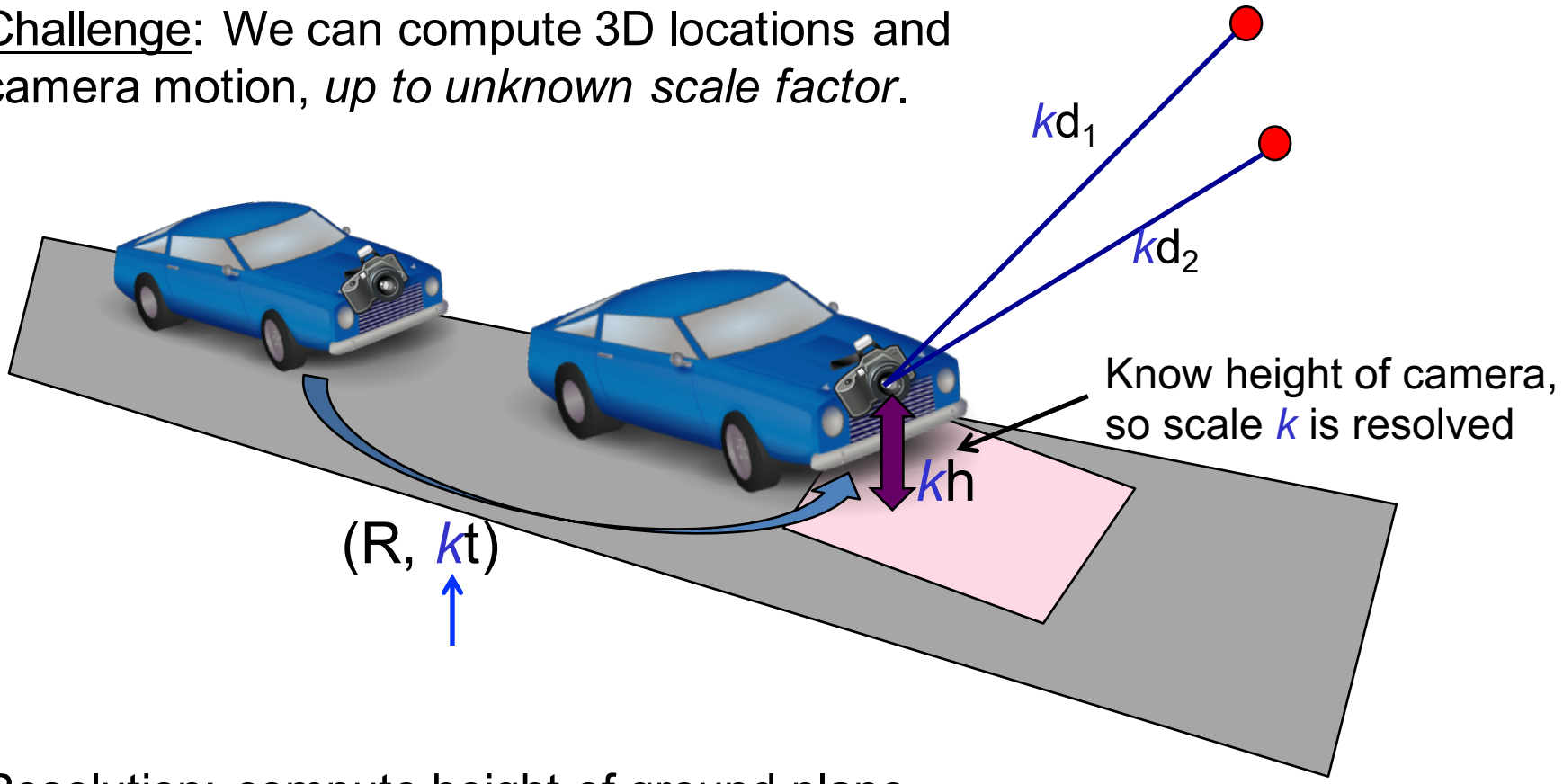


Every bundle adjustment causes 3D points and cameras to change



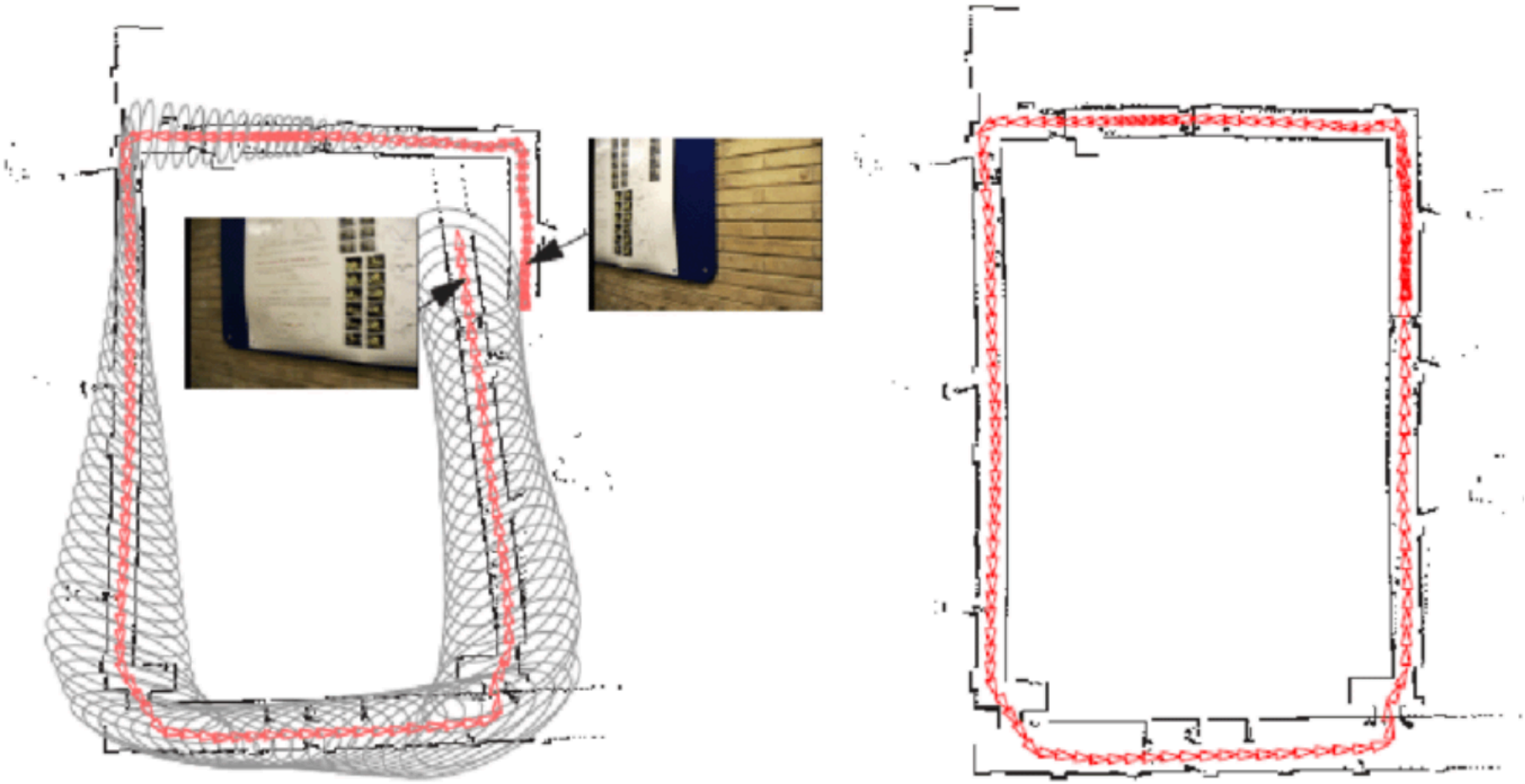
Scale Drift Correction

Challenge: We can compute 3D locations and camera motion, *up to unknown scale factor*.



Resolution: compute height of ground plane.

Loop Closure



Recognize whether same image is observed in some previous frame.

Some Recipes for SFM to Work

- Do everything you can to remove outliers

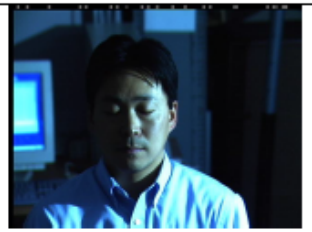
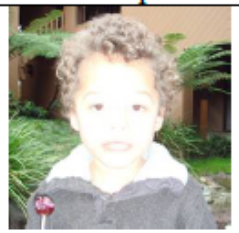
- Solve minimal problems to estimate geometric entities
  - Keeps RANSAC tractable
  - Typically, expect to spend 0.01ms

- Strategically consider what variables to optimize
  - Keyframe-based designs are successful
  - Try to robustly build long feature tracks
  - Do bundle adjustment whenever possible

- Drift is inevitable, so have a plan to address it
  - Local scale correction and global pose correction when possible

Module 2: Faces and Humans

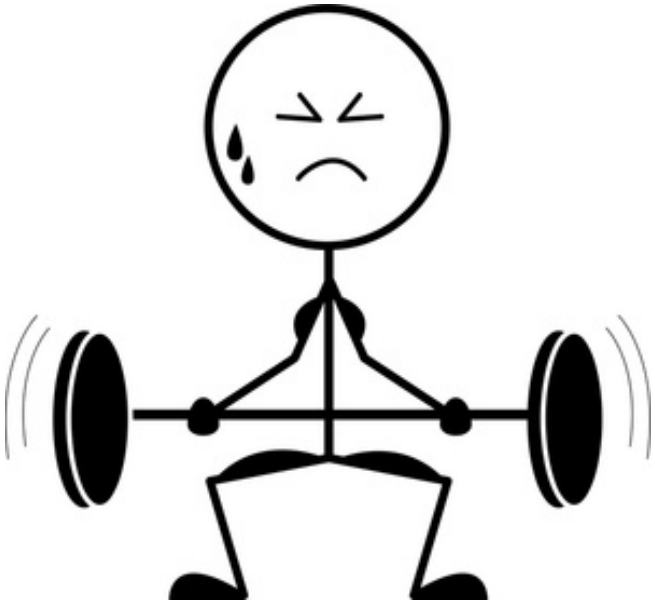
Unconstrained Face Recognition

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%



Face Recognition on LFW Benchmark

- Human performance : **99.20%**
- Local Binary Patterns : 95.17%
- DeepFace : 97.35 %

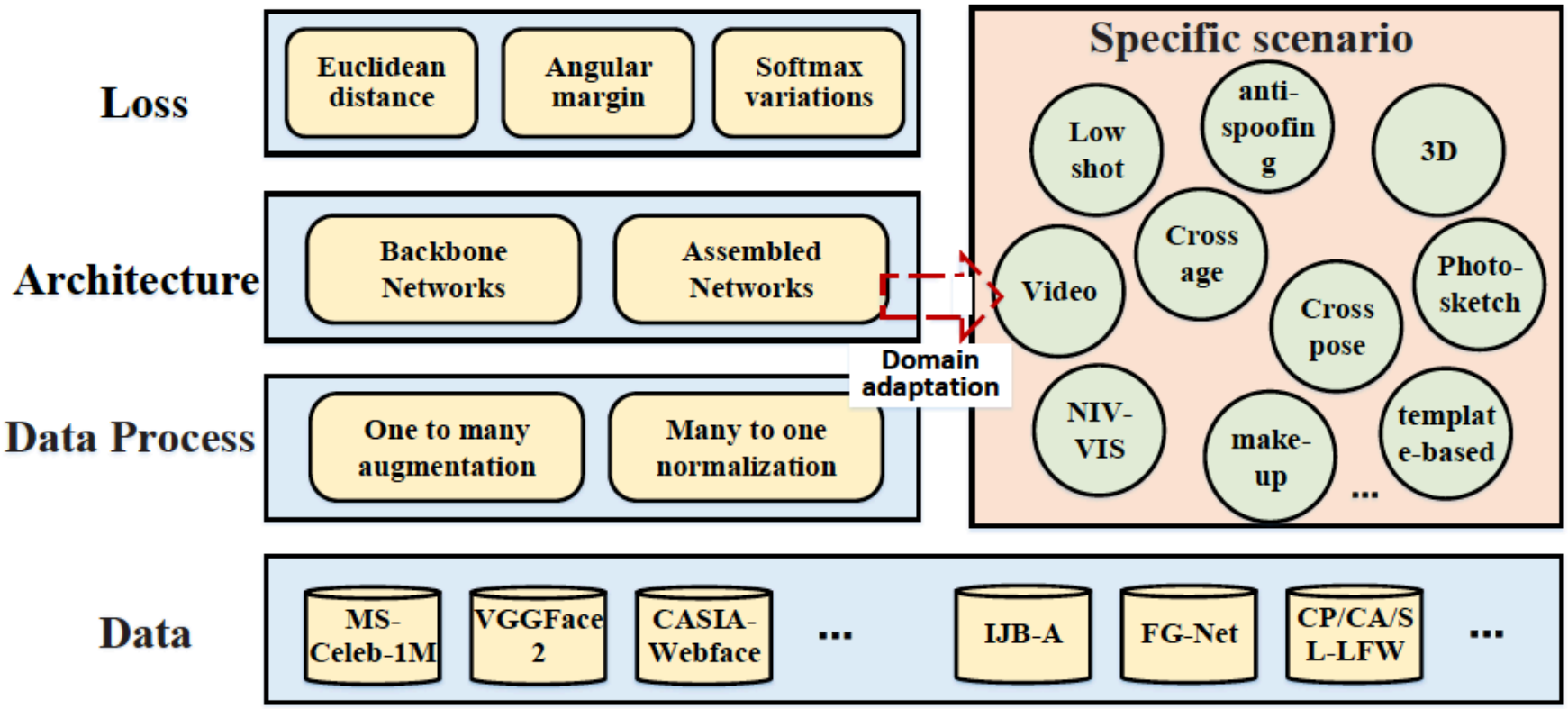
Face Recognition on LFW Benchmark

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

Face Recognition on LFW Benchmark

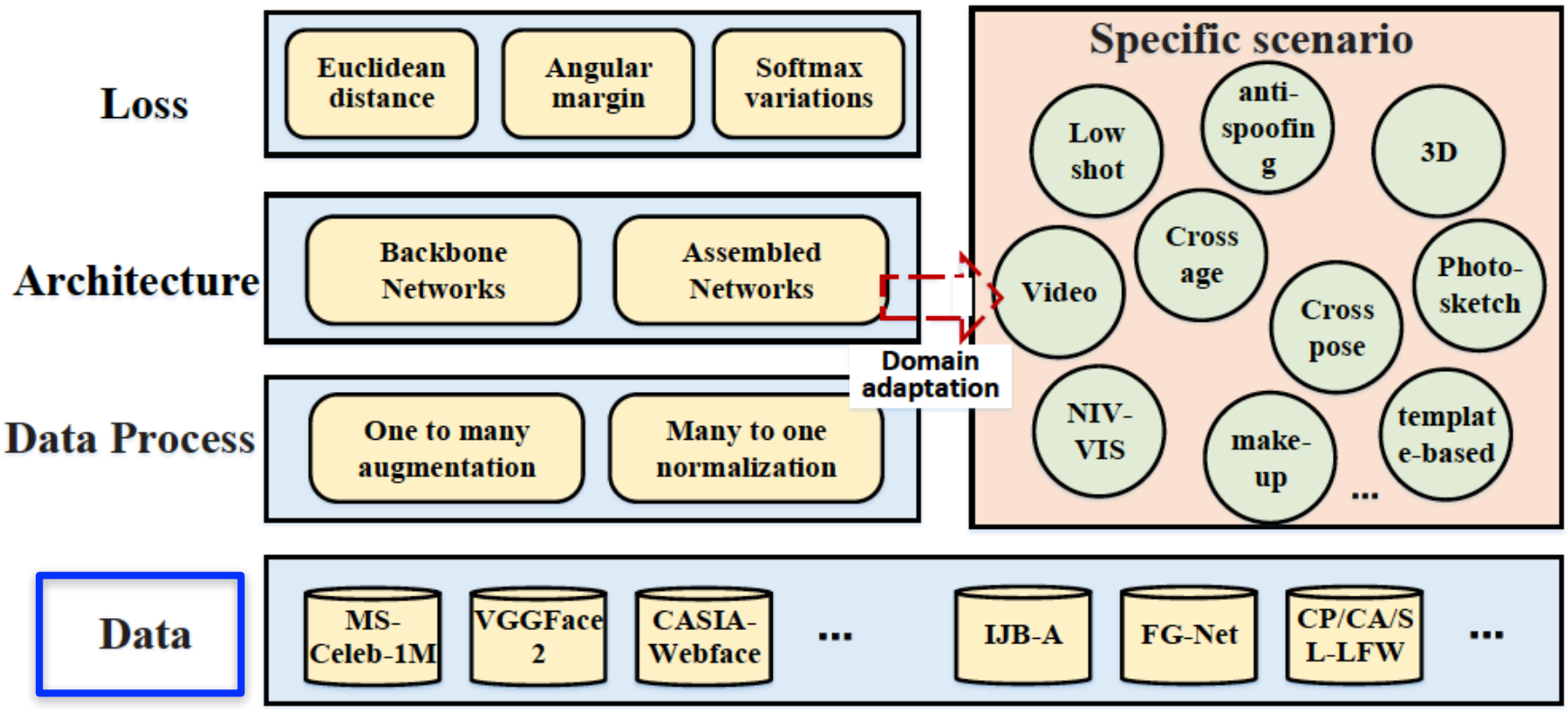
- Human performance : **99.20%**
- Local Binary Patterns : 95.17%
- DeepFace : 97.35 %
- DeepID2 : 99.15%
- FaceNet : **99.63%**

Axes for Studying Face Recognition



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

Axes for Studying Face Recognition



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

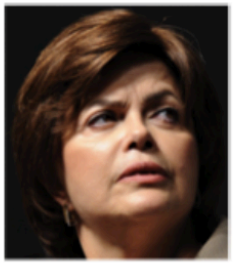
Labeled Faces in the Wild (LFW) Dataset

- Face Verification:
  - Specify whether pair of images belong to the same person
- 13K images, 5.7K people
- Standard benchmark in the community
- Test protocols depending upon availability of outside training data

[G. Huang, M. Ramesh, T. Berg and E. Learned-Miller]

IJB-A Dataset

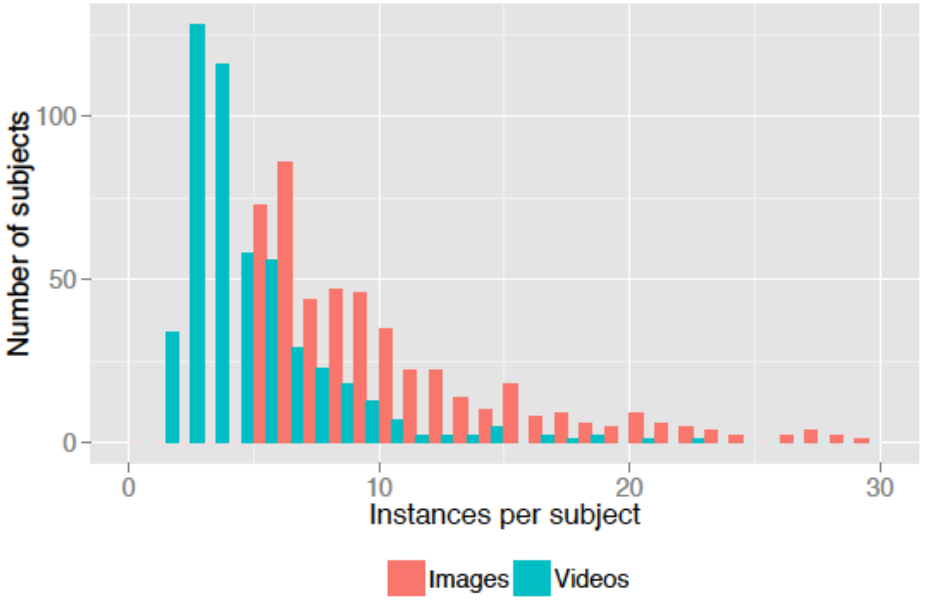
(a)



(b)

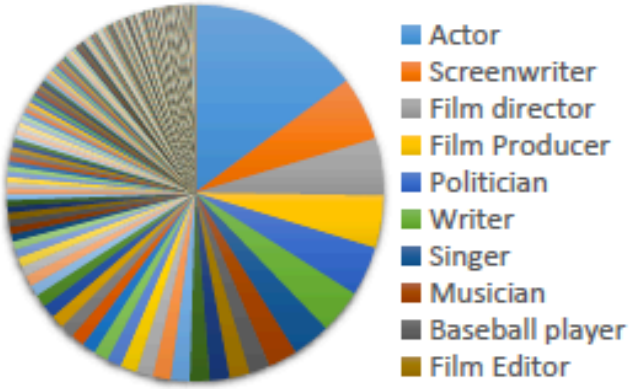
(c)

	(a)	(b)	(c)
Types of Images:	Frontal, Cooperative subject, Controlled environment	Near frontal, uncooperative, minimal environment variations (e.g., LFW)	Full variation in pose, illumination, environment
Automated detection ability:	Human performance	Near human performance	Cannot detect consistently
Automated recognition ability:	Human performance	Near human performance	Cannot recognize

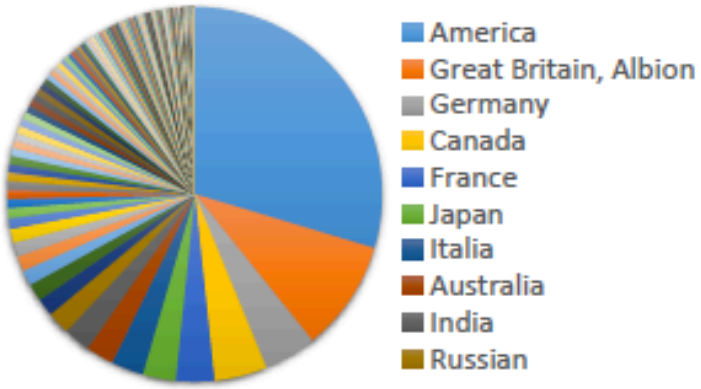


[Klare et al., CVPR 2015]

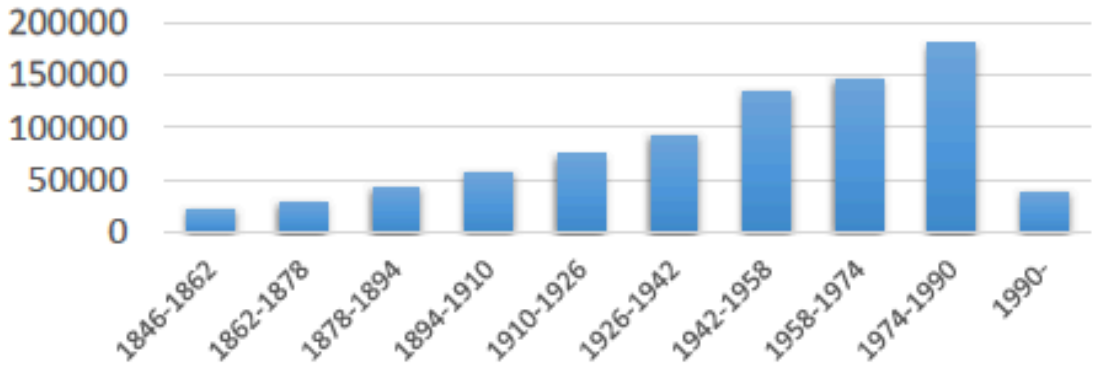
MS-1M Dataset



(a) Professions



(b) Nationality

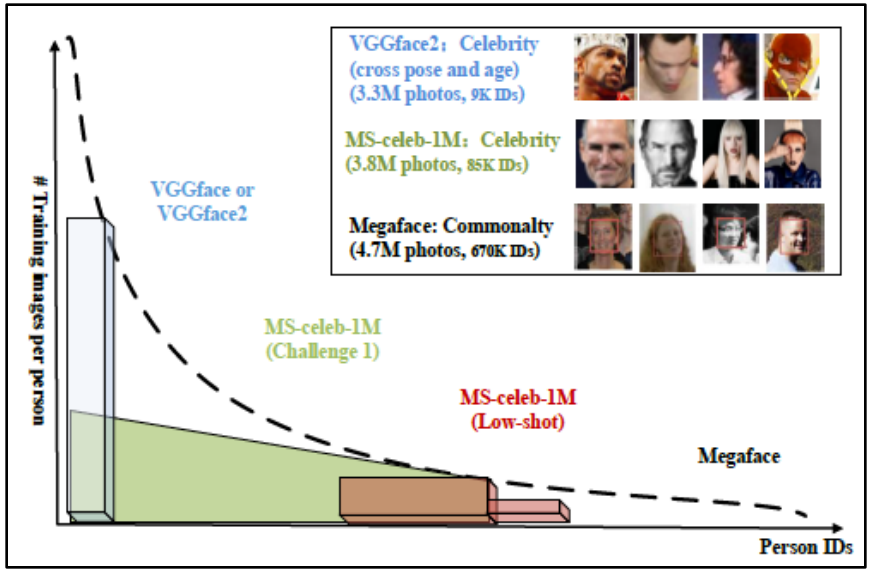


(c) Date of Birth

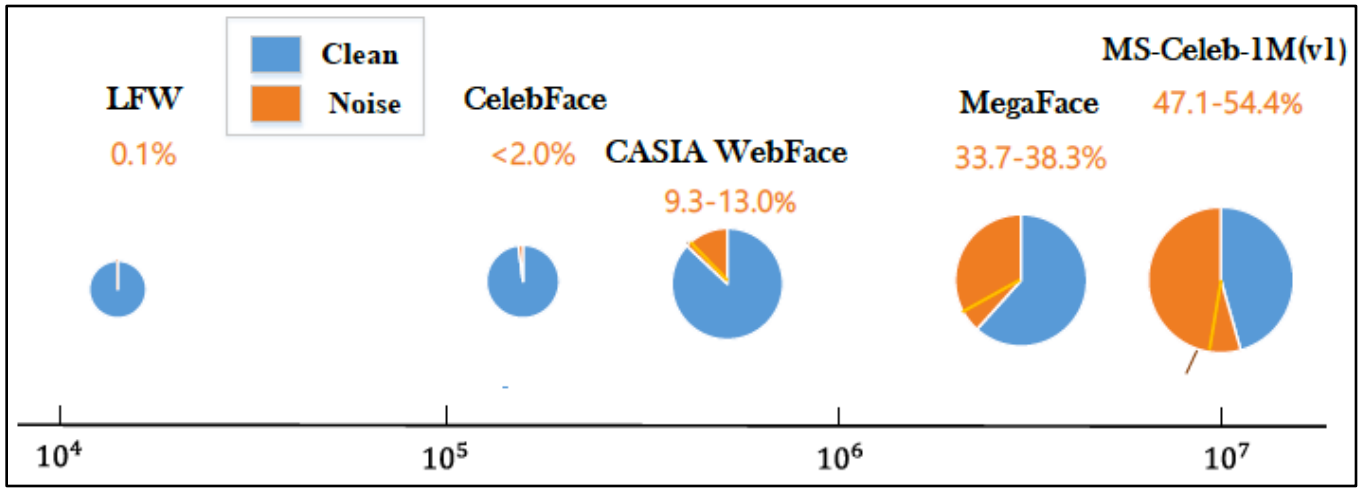
(d) Gender

[Guo et al., ECCV 2016]

Face Datasets



Variation in number of faces per identity

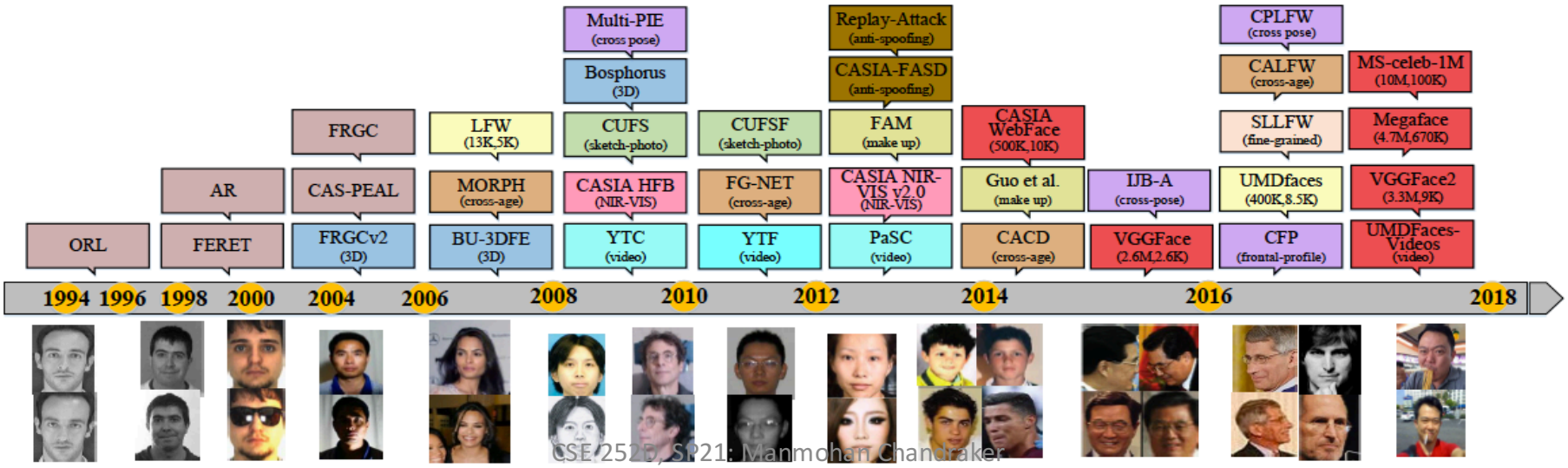


Variation in approximate labeling noise levels

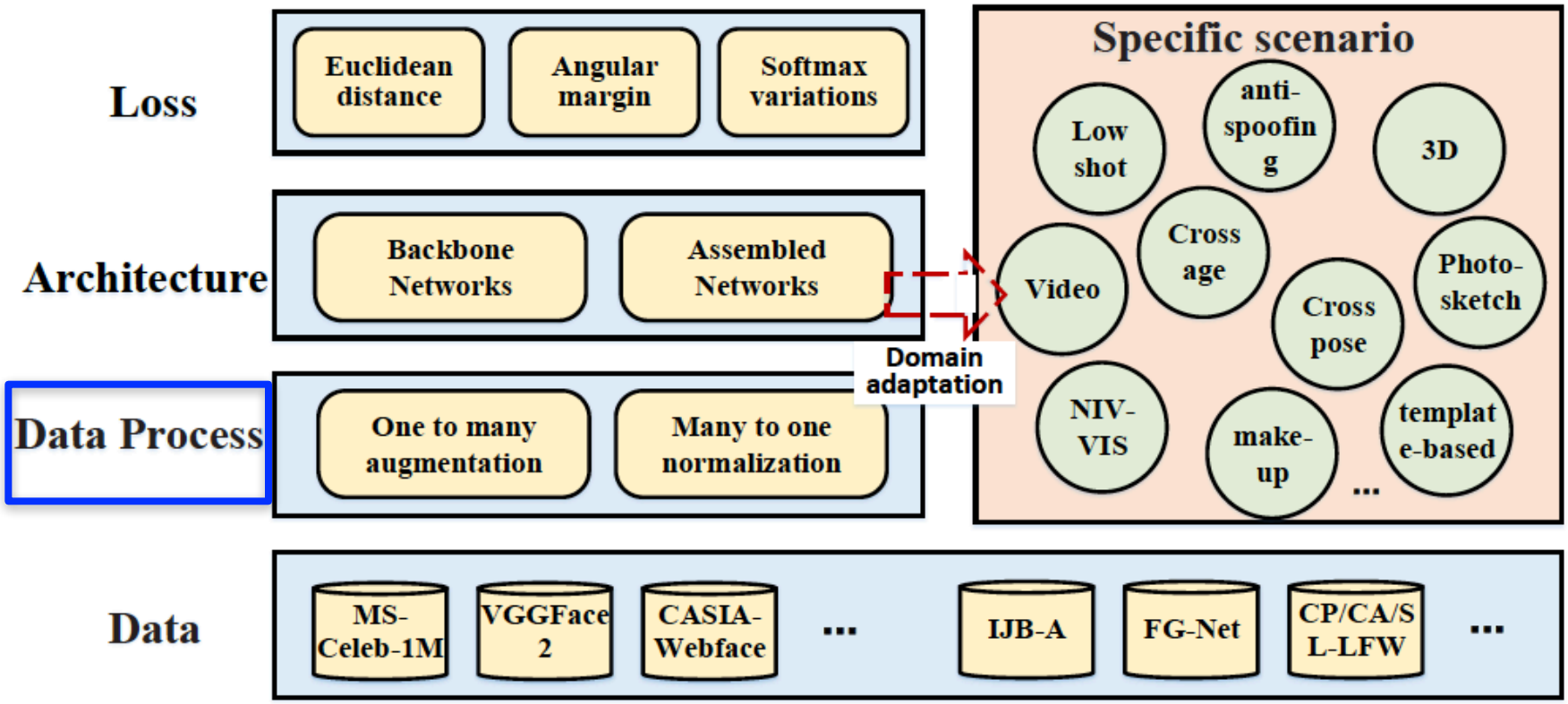
[Guo et al., ECCV 2016]

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

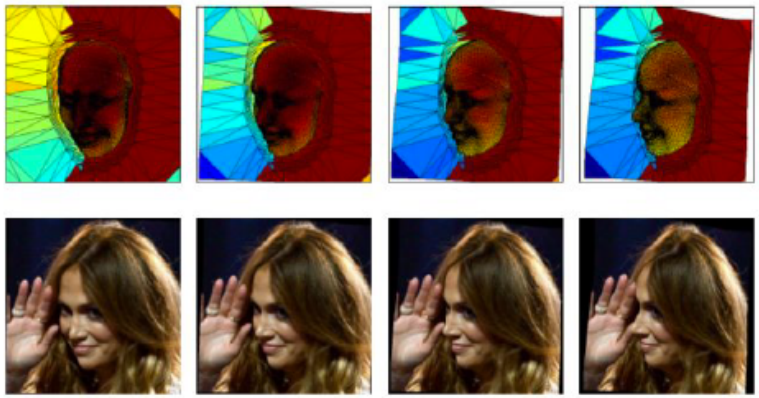


Axes for Studying Face Recognition



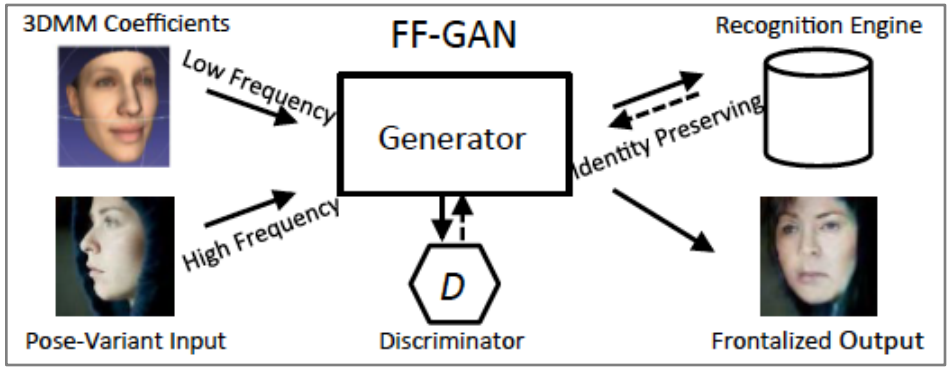
Face Processing

Augmentation



[Zhu et al., "3DDFA"]

Normalization

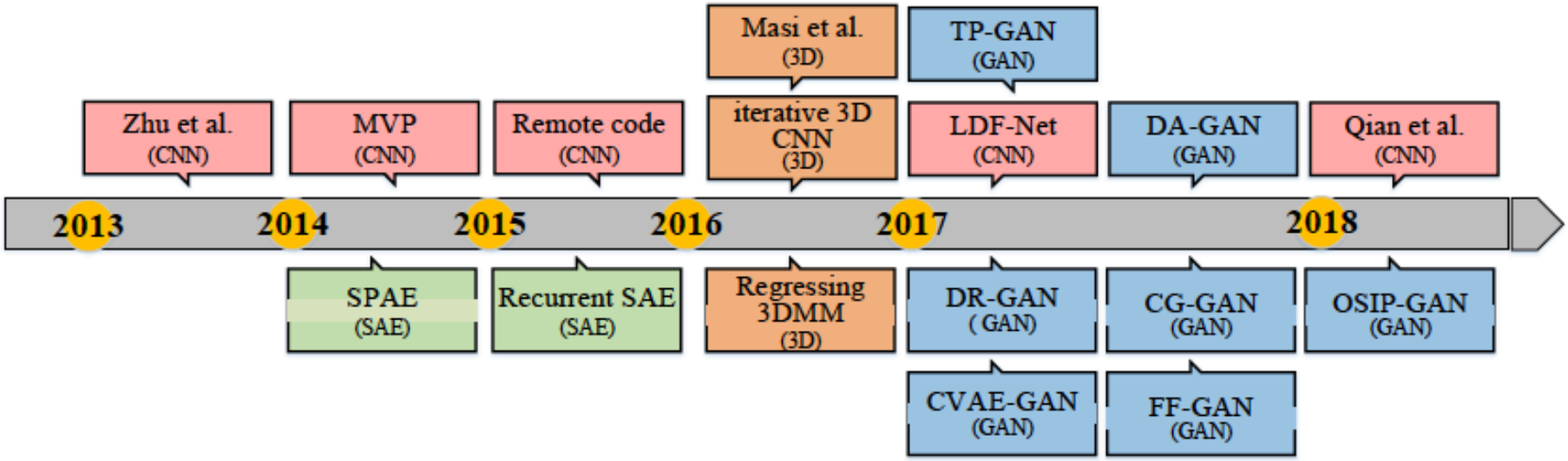


[Yin et al., "FF-GAN"]

- **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

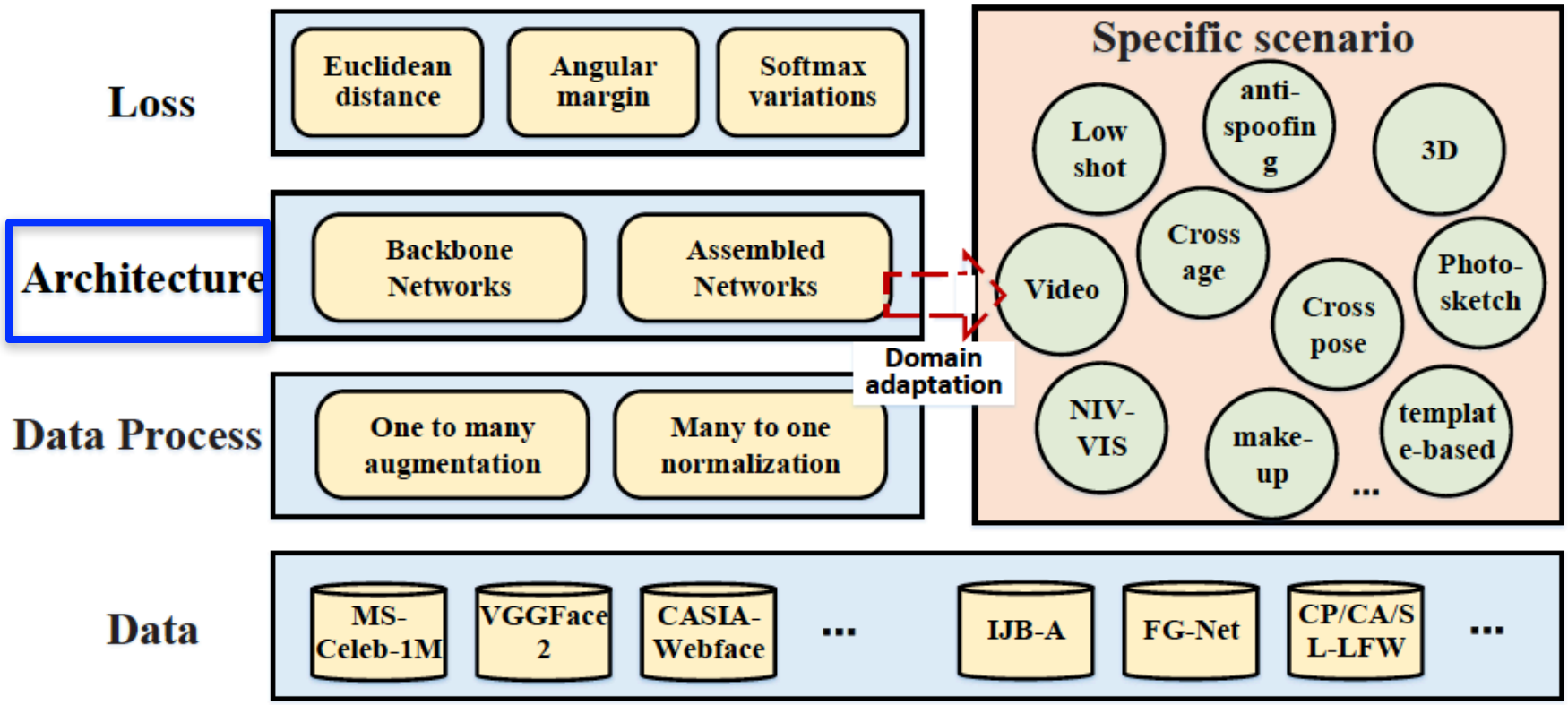
Face Processing



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

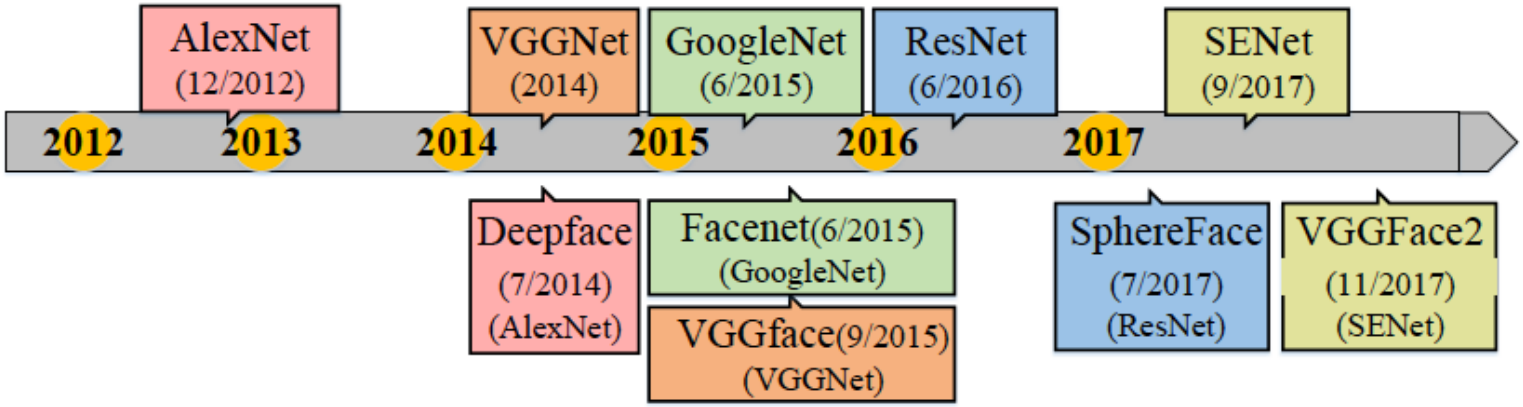
- **Many-to-One Normalization:** reduce variation in test-time inputs
  - Generate frontal face from pose-variant input
  - Use GANs or methods to generate faces with neutral attributes

Axes for Studying Face Recognition

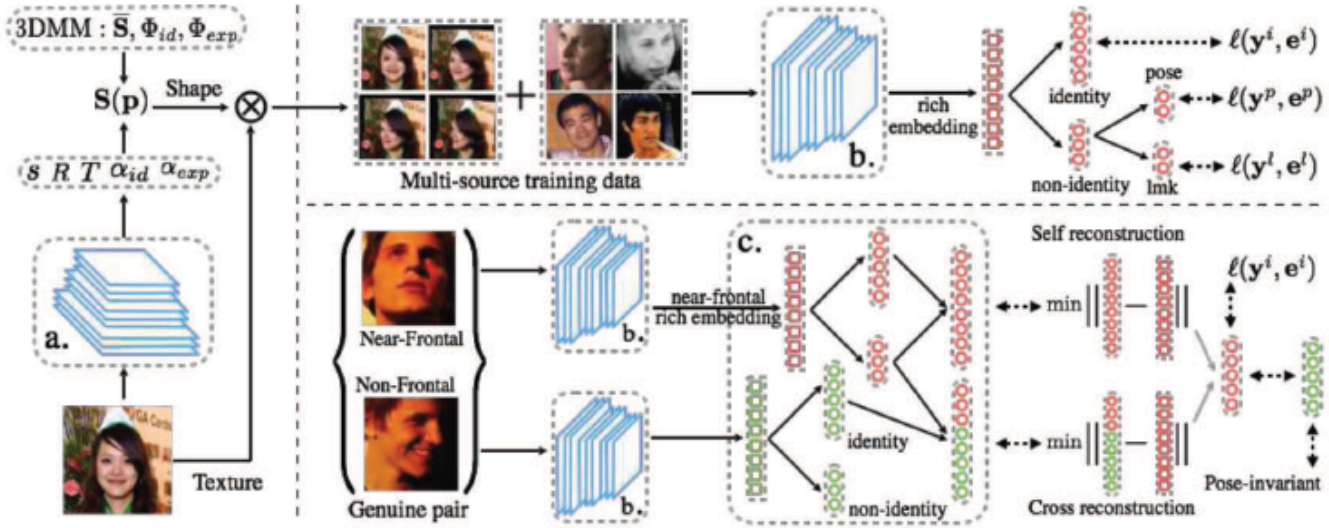


Network Architectures

Backbone networks

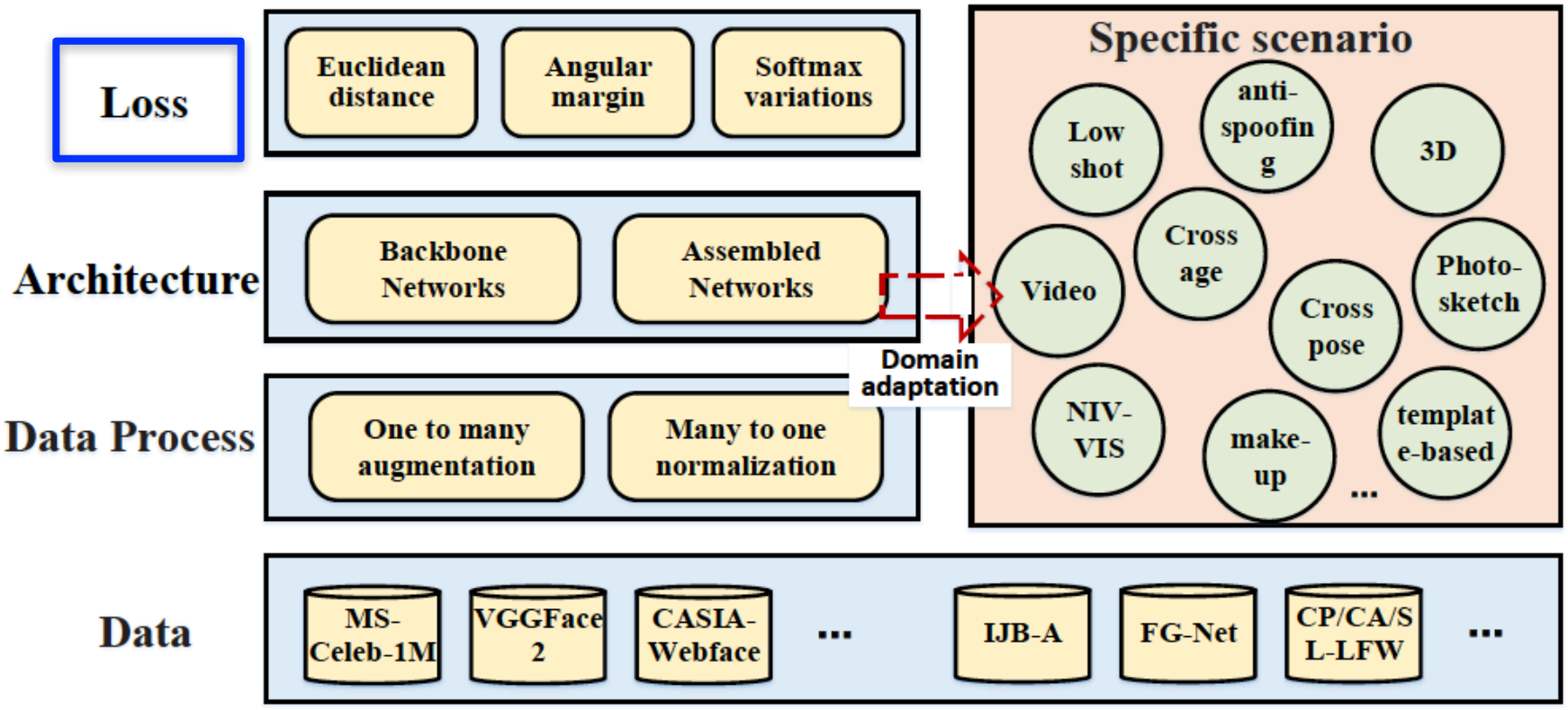


Multi-tasked networks



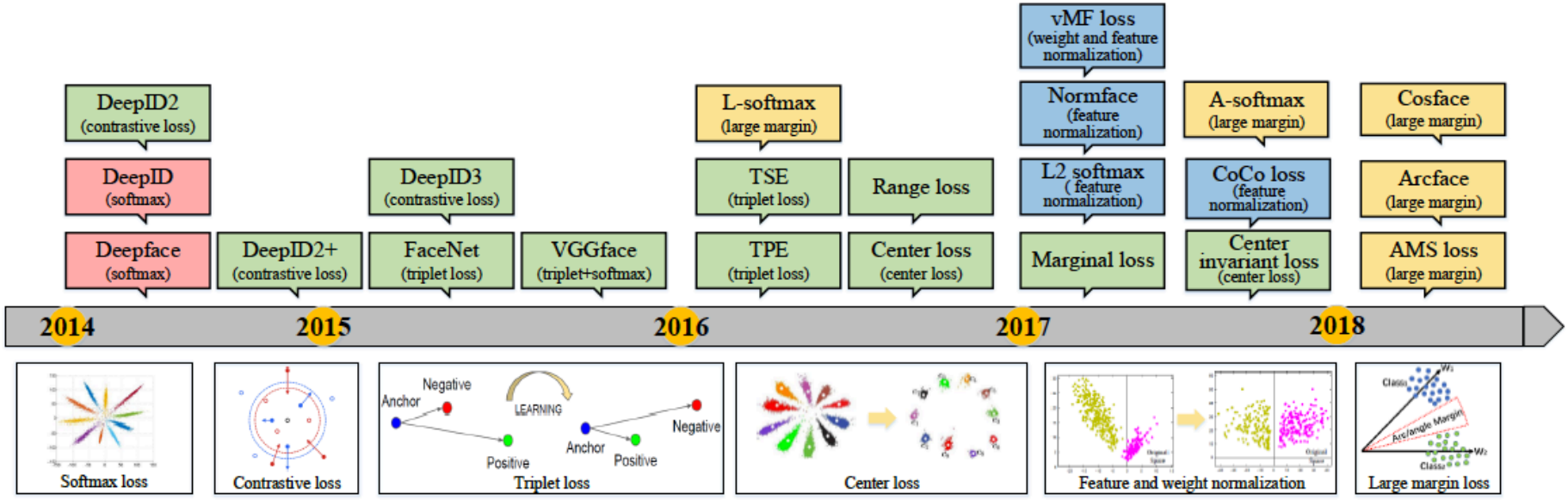
[Peng et al., ICCV 2017]

Axes for Studying Face Recognition



Loss Functions

Large-margin losses and softmax variants



Performance of Various Loss Functions

Method	Public. Time	Loss	Architecture	Number of Networks	Training Set	Accuracy±Std(%)
DeepFace [195]	2014	softmax	Alexnet	3	Facebook (4.4M,4K)	97.35±0.25
DeepID2 [187]	2014	contrastive loss	Alexnet	25	CelebFaces+ (0.2M,10K)	99.15±0.13
DeepID3 [188]	2015	contrastive loss	VGGNet-10	50	CelebFaces+ (0.2M,10K)	99.53±0.10
FaceNet [176]	2015	triplet loss	GoogleNet-24	1	Google (500M,10M)	99.63±0.09
Baidu [124]	2015	triplet loss	CNN-9	10	Baidu (1.2M,18K)	99.77
VGGface [149]	2015	triplet loss	VGGNet-16	1	VGGface (2.6M,2.6K)	98.95
light-CNN [225]	2015	softmax	light CNN	1	MS-Celeb-1M (8.4M,100K)	98.8
Center Loss [218]	2016	center loss	Lenet+-7	1	CASIA-WebFace, CACD2000, Celebrity+ (0.7M,17K)	99.28
L-softmax [126]	2016	L-softmax	VGGNet-18	1	CASIA-WebFace (0.49M,10K)	98.71
Range Loss [261]	2016	range loss	VGGNet-16	1	MS-Celeb-1M, CASIA-WebFace (5M,100K)	99.52
L2-softmax [157]	2017	L2-softmax	ResNet-101	1	MS-Celeb-1M (3.7M,58K)	99.78
Normface [206]	2017	contrastive loss	ResNet-28	1	CASIA-WebFace (0.49M,10K)	99.19
CoCo loss [130]	2017	CoCo loss	-	1	MS-Celeb-1M (3M,80K)	99.86
vMF loss [75]	2017	vMF loss	ResNet-27	1	MS-Celeb-1M (4.6M,60K)	99.58
Marginal Loss [43]	2017	marginal loss	ResNet-27	1	MS-Celeb-1M (4M,80K)	99.48
SphereFace [125]	2017	A-softmax	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.42
CCL [155]	2018	center invariant loss	ResNet-27	1	CASIA-WebFace (0.49M,10K)	99.12
AMS loss [205]	2018	AMS loss	ResNet-20	1	CASIA-WebFace (0.49M,10K)	99.12
Cosface [207]	2018	cosface	ResNet-64	1	CASIA-WebFace (0.49M,10K)	99.33
Arcface [42]	2018	arcface	ResNet-100	1	MS-Celeb-1M (3.8M,85K)	99.83
Ring loss [272]	2018	Ring loss	ResNet-64	1	MS-Celeb-1M (3.5M,31K)	99.50

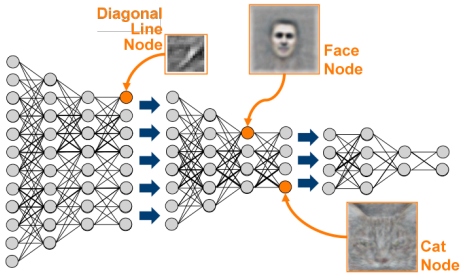
Learning Face Representations

Steps in Face Recognition

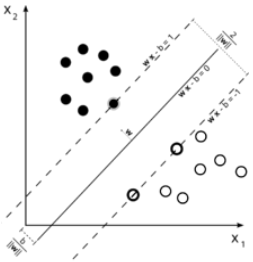
- Face Detection
  - Localize the face

- Face Alignment
  - Factor out 3D transformation

- Feature Extraction
  - Find compact representation



- Classification
  - Answer the question



Challenges in Face Alignment

- Infer 3D from 2D
  - Slight occlusion
  - Lighting condition
  - Head orientation
  - Non rigid deformation

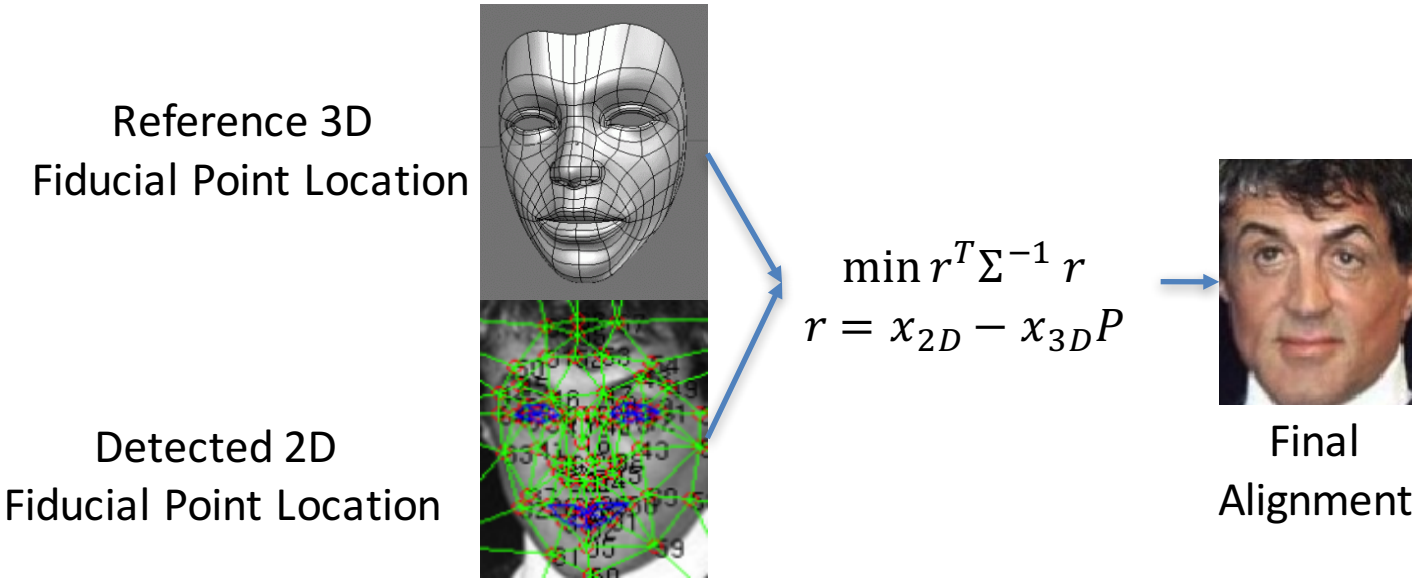
Face Alignment: Substep 1

- 2D feature point extraction
- 2D alignment $x_{anchor} = (S * R * T)x_{source}$
- Only for in plane alignment

[Taigman et al., DeepFace, CVPR 2014]

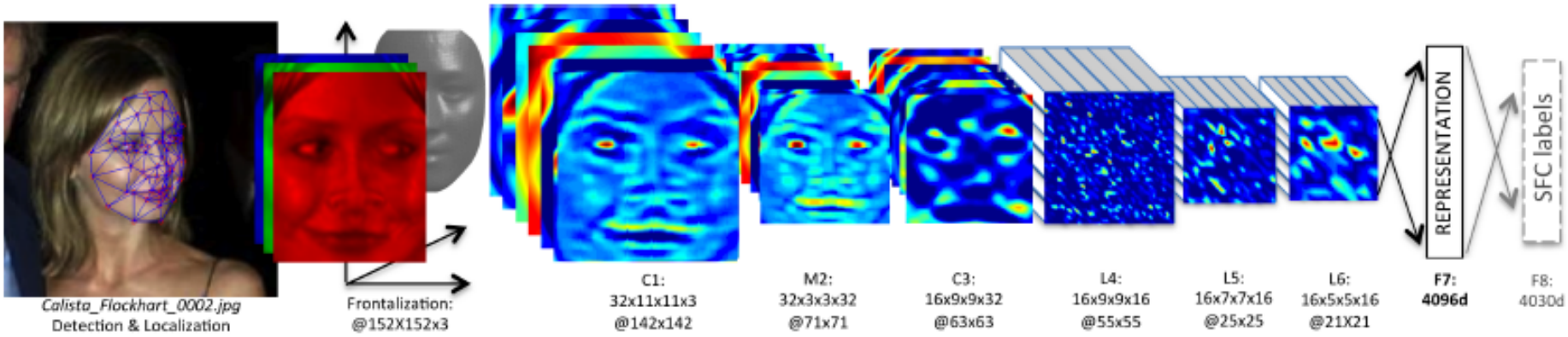
Face Alignment: Substep 2

- 3D feature point extraction
- 3D alignment: piecewise affine transformation



[Taigman et al., DeepFace, CVPR 2014]

Architecture

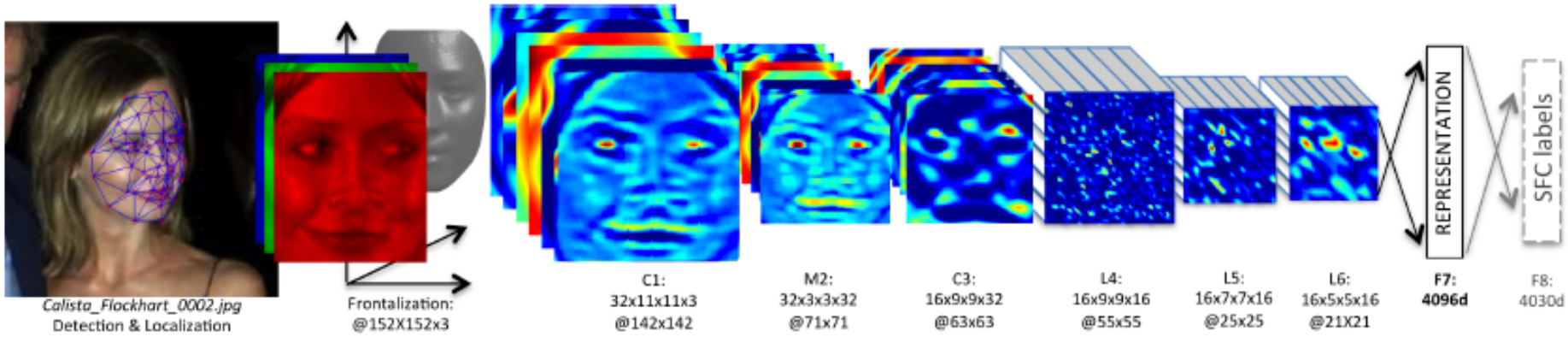


Layer 1-3 : Intuition

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

[Taigman et al., DeepFace, CVPR 2014]

Architecture



Layer 1-3 : Intuition

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

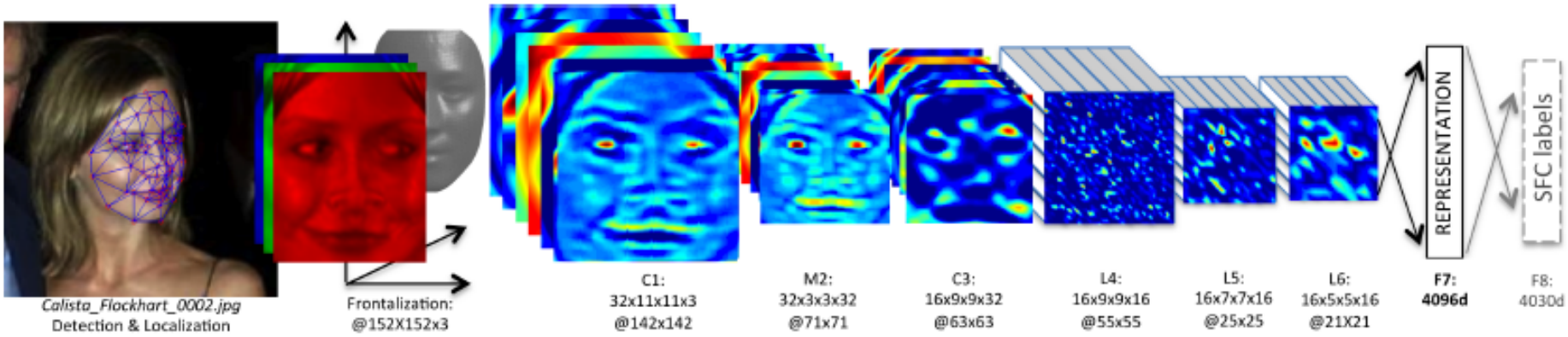
Layer 4-6: Intuition

- 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]

Architecture



- Layer F7 is fully connected and generates 4096d vector
- Sparse representation of face descriptor
- Layer F8 is fully connected and generates 4030d vector
- F8 calculates probability with softmax $p_k = \frac{\exp(o_k)}{\sum_h \exp(o_h)}$
- Cross-entropy loss function: $\sum_{i=1}^{n} -p_i \log \hat{p}_i$

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

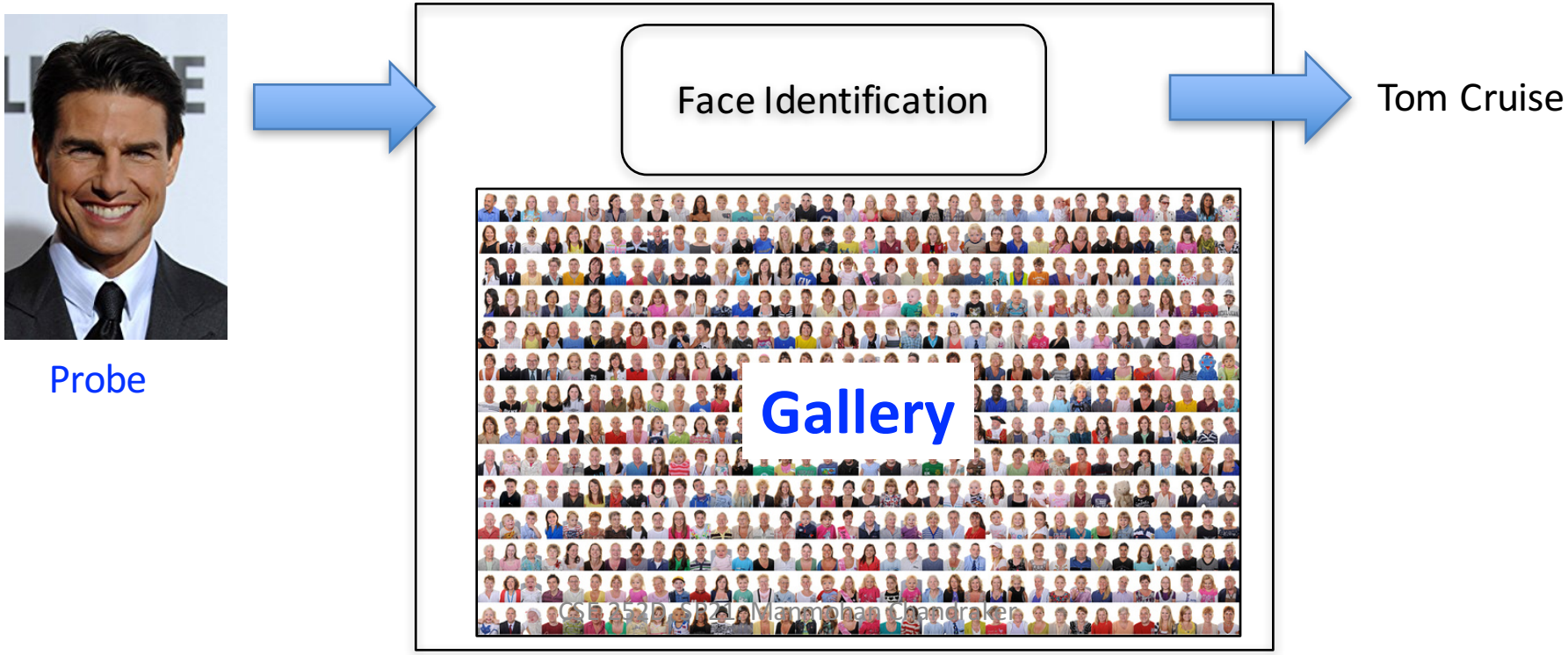
Predicted probability distribution

[Taigman et al., DeepFace, CVPR 2014]

Face Identification

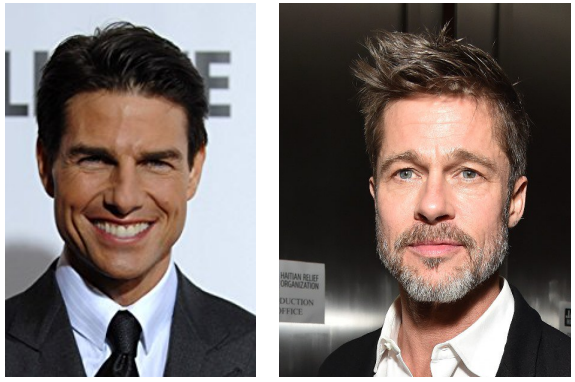
- Closed set identification: assign one of gallery identities to probe image
- Galleries can be very large, high chance of similar appearances

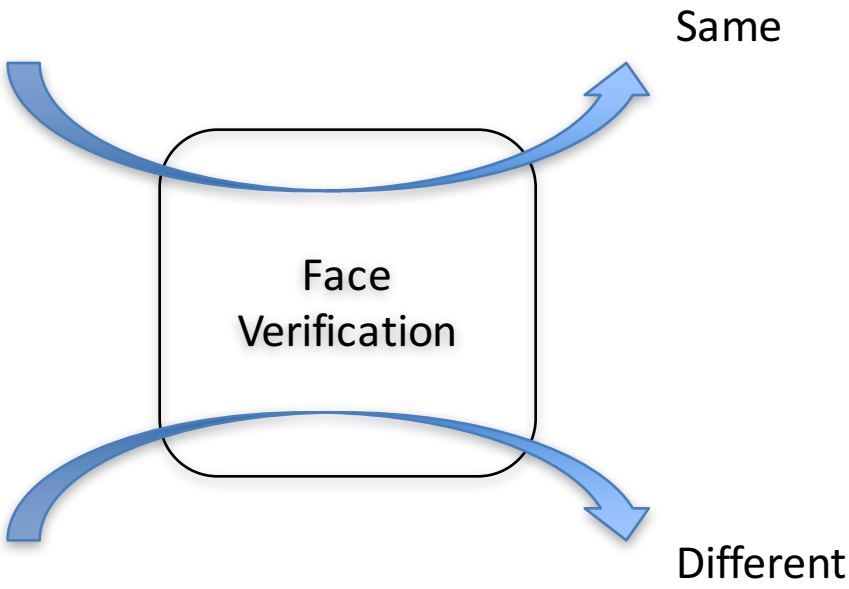
- Goal is to have sharp decision boundary between gallery identities
- Feature need not generalize to other tasks (identities outside the gallery)



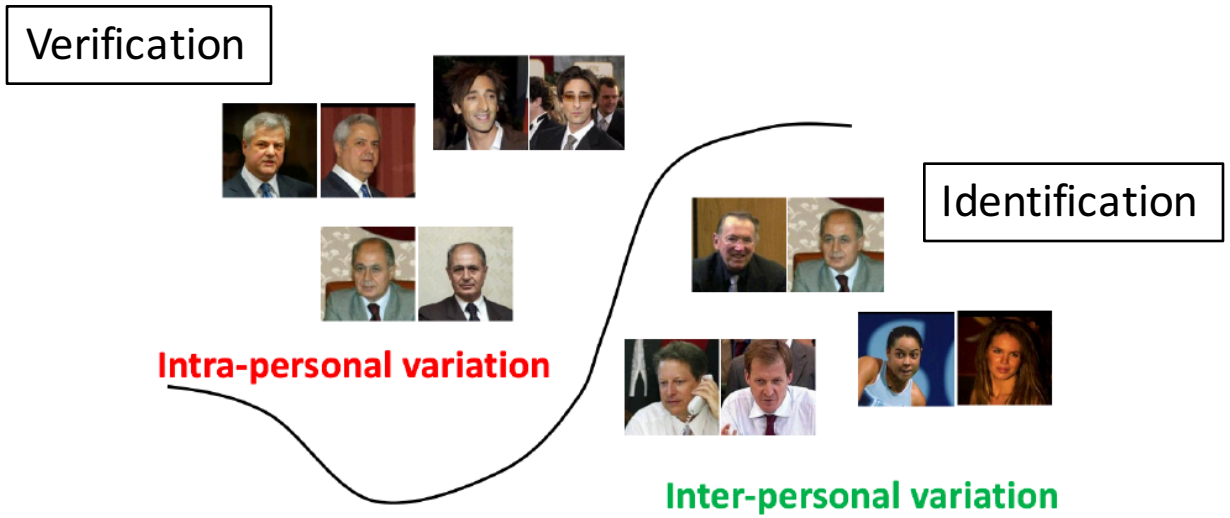
Face Verification

- 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 $D(x_i, x_j)$





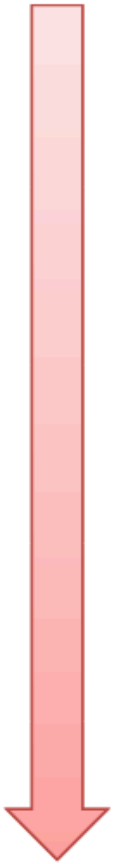
Verification and Identification Signals



- **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

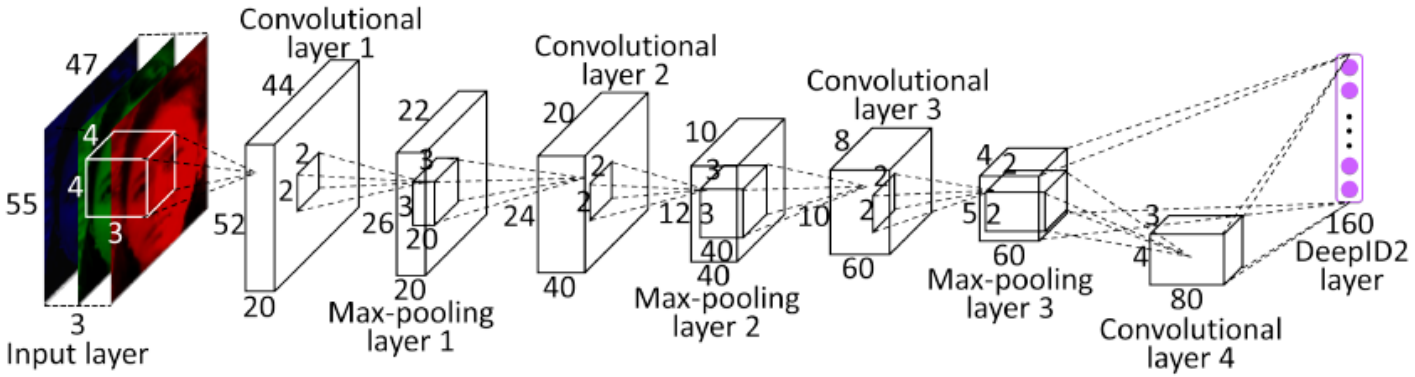
Predicting multi-class labels (identification)

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

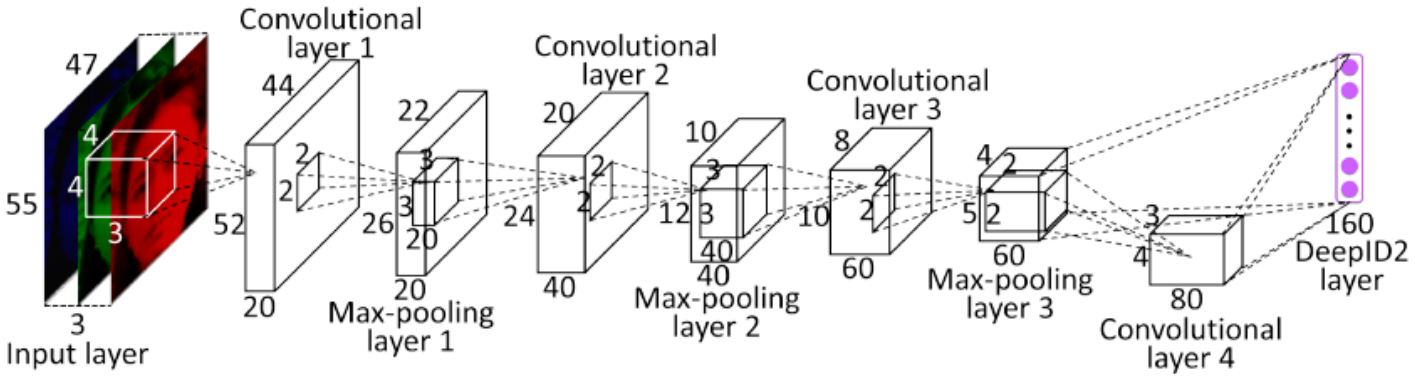
Target probability distribution
p_i = 1 for class t and 0 for other i

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:

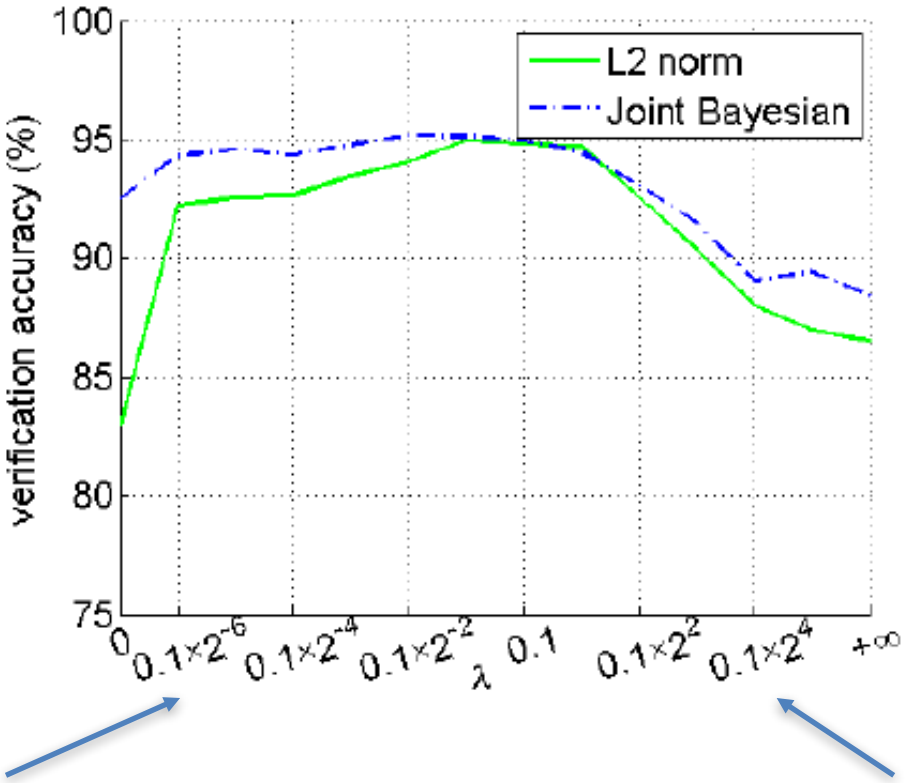
$$\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \frac{1}{2}\left(y_{ij} - \sigma(wd + b)\right)^2 \text{ , binary } y_{ij}, \quad d = \frac{f_i \cdot f_j}{\|f_i\|_2 \|f_j\|_2}$$

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

 [Sun et al., DeepID2, NeurIPS 2014]

Balancing Identification and Verification

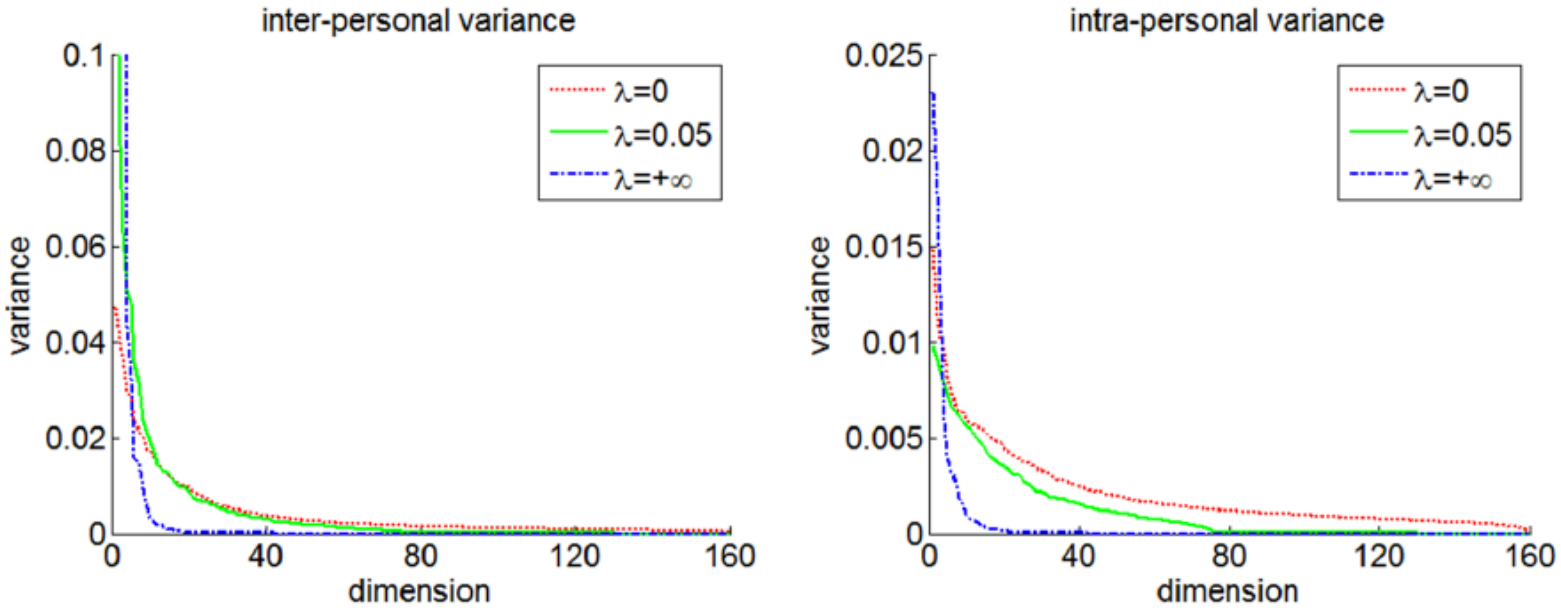
- Balance required between signals to learn good features
- High λ : more weight on verification



Small λ : only identification

Large λ : only verification

Balancing Identification and Verification



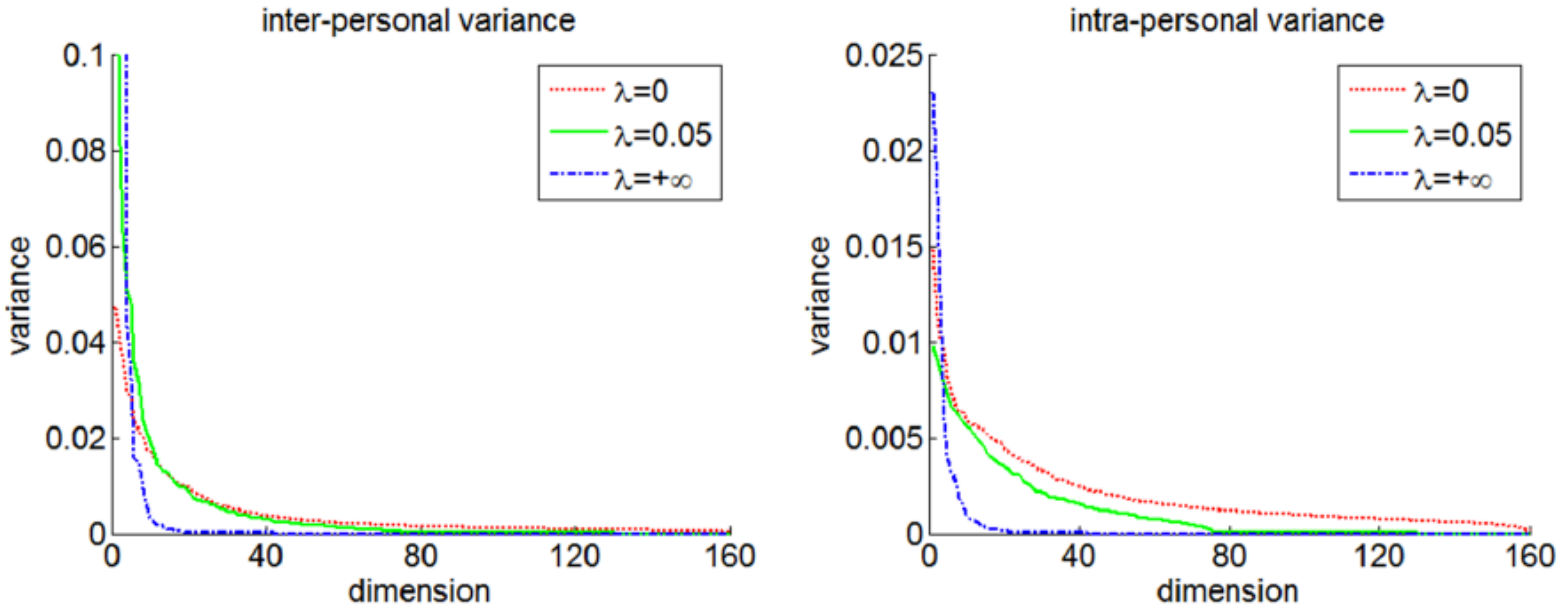
- Inter-class scatter : $\sum_{i=1}^{c} n_i \cdot (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^\top$

c classes

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

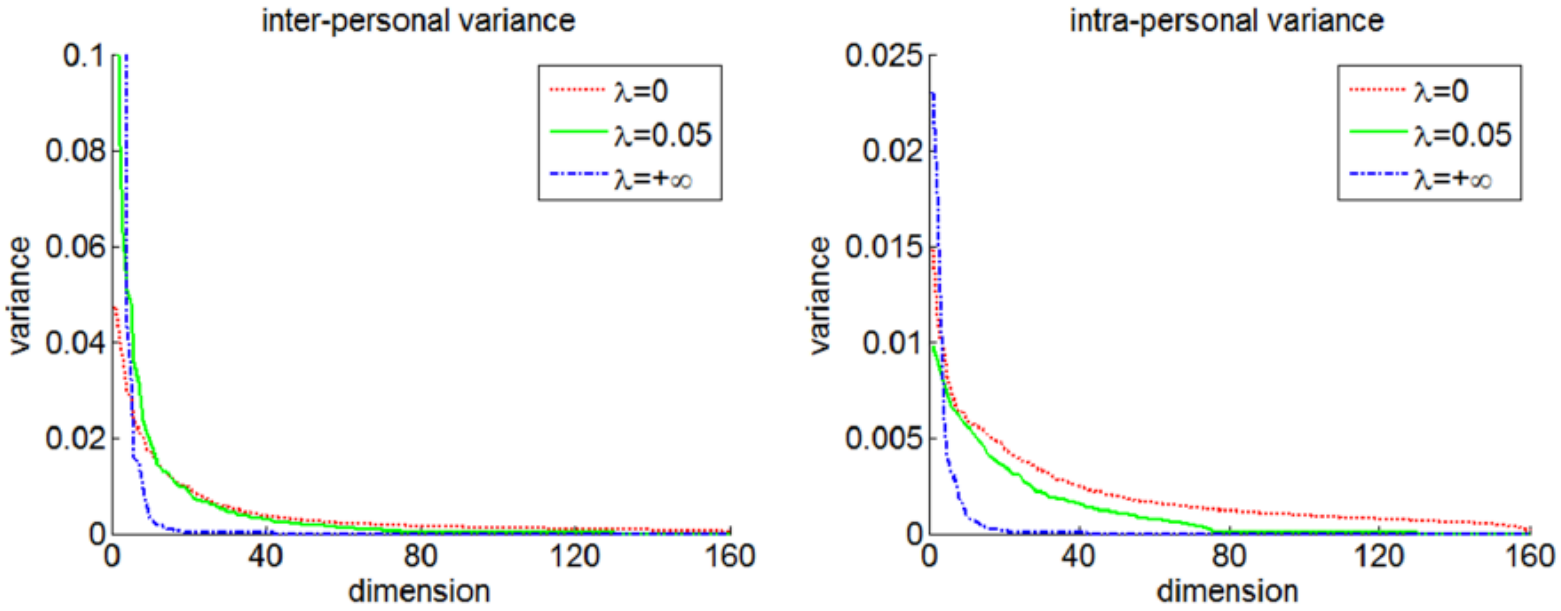
- Variance in scatter indicated by size of eigenvalues
- Small number of eigenvectors: diversity of variation is low
- Both diversity and magnitude of feature variance matters for recognition

Balancing Identification and Verification



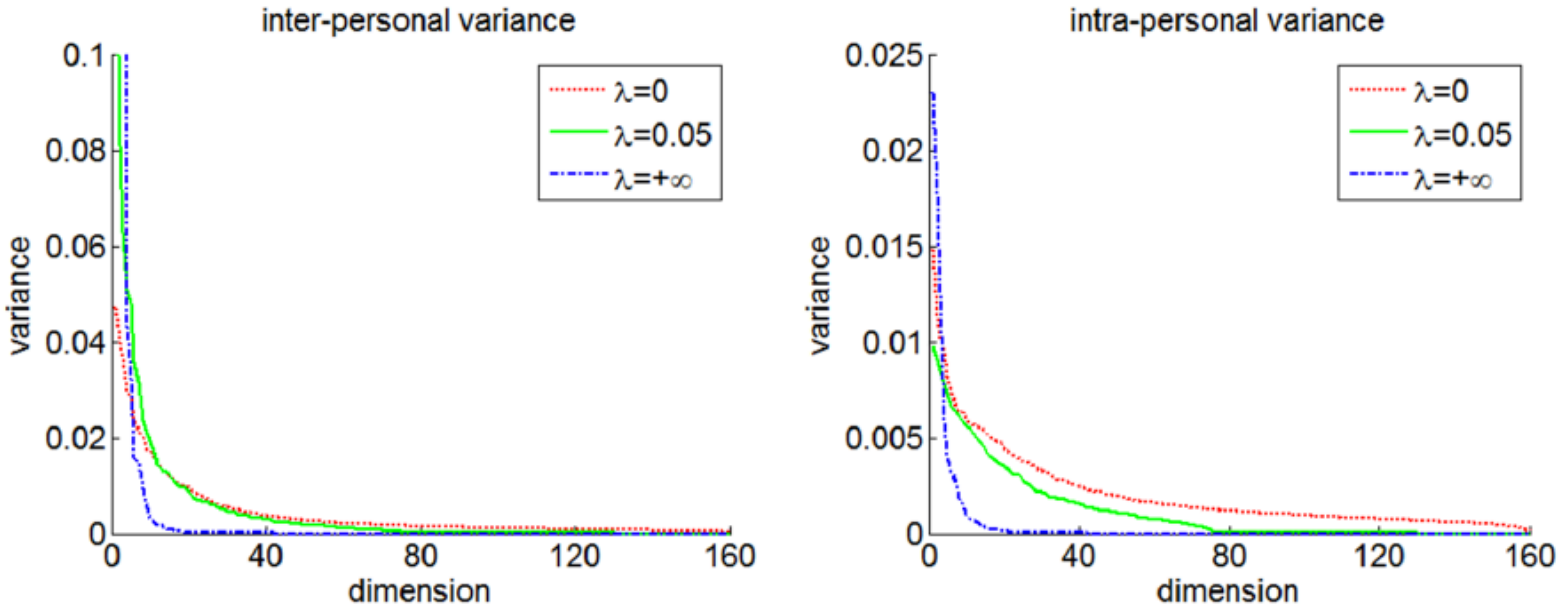
- When only identification signal is used $(\lambda = 0)$:
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification

Balancing Identification and Verification



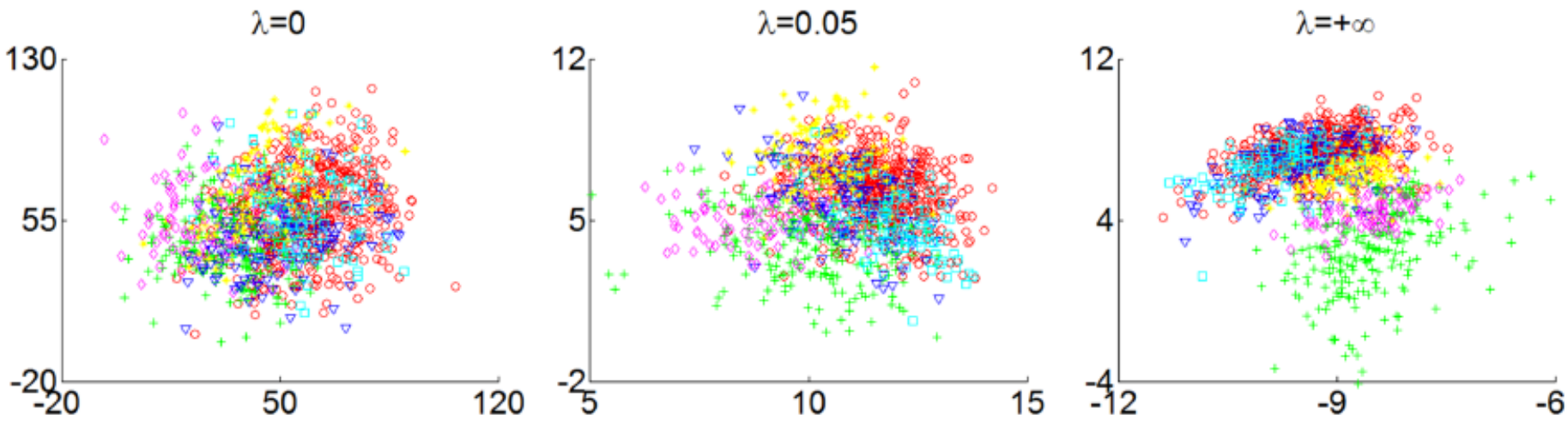
- When only identification signal is used $(\lambda = 0)$:
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification
- When only verification signal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