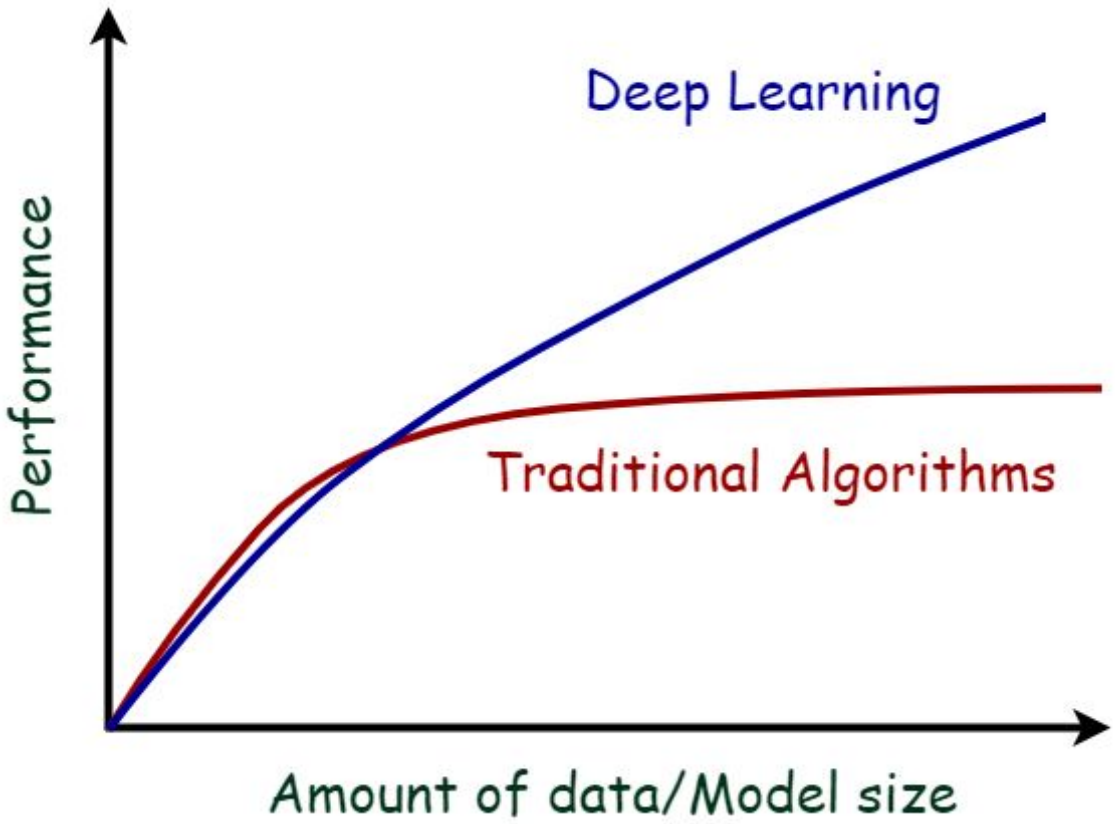
A Survey on Deep Learning based Face Recognition

Na Zhang
Dr. Guodong Guo
LCSEE, WVU
Aug 2019

- Overview
- Deep Learning Methods on FR
- Specific Face Recognition Problems
- Face Databases
- Discussion and Challenges

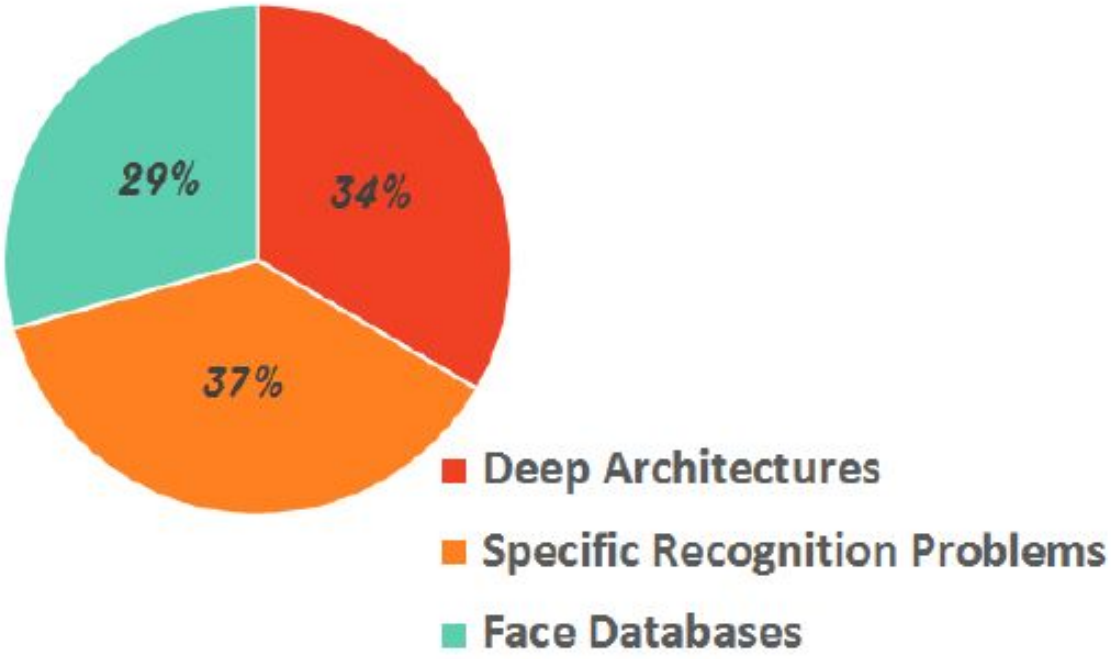
Deep Learning

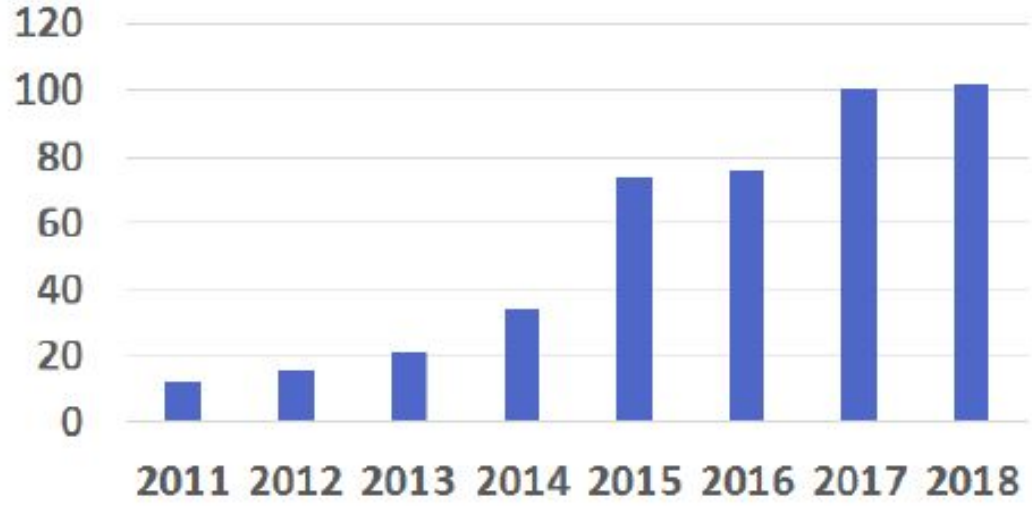
- Recently, Deep Neural Network has established itself as a dominant technique in machine learning
- Deep and large networks have exhibited impressive results when there are large training data sets and computation resources (CPU cores and/or GPUs)



- This survey presents a comprehensive overview of about 330 face recognition works using deep learning within the recent years
- It shows that:
  - ✔ DL has been fully applied to FR and plays important roles;
  - ✔ Many specific issues or challenges have been addressed in FR by DL, e.g., pose, illumination, expression, 3D, heterogenous matching;
  - ✔ Various face datasets have been collected in recent years, including still images, videos, and heterogeneous data.

Paper distribution

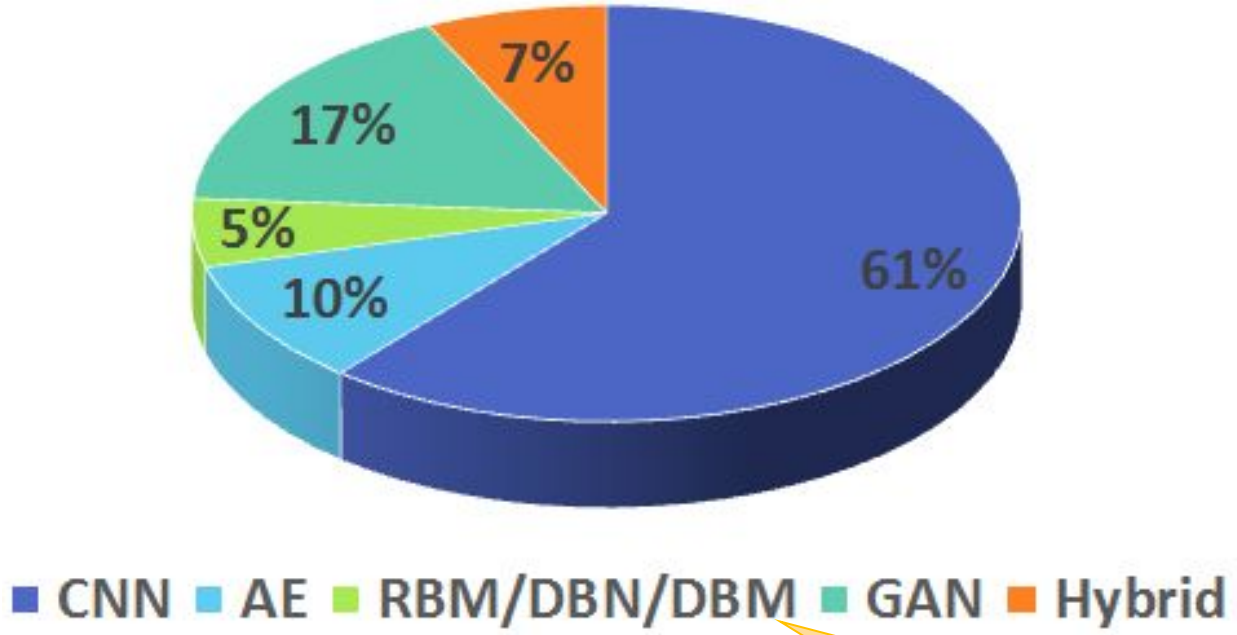




- Overview

- Deep Learning Methods on FR
- Specific Face Recognition Problems
- Databases
- Discussions and Challenges

- Many types of neural networks were used in FR



- ✔ CNN: Convolutional Neural Network
- ✔ AE: Autoencoder
- ✔ GAN: Generative Adversarial Network
- ✔ DBM: Deep Boltzmann Machine
- ✔ DBN: Deep Belief Network
- ✔ RBM: Restricted Boltzmann Machine

CNN: the most popular
AE: gained much attention
GAN: gains increasing attention recently

❖ CNN based deep methods

❑ Single CNN

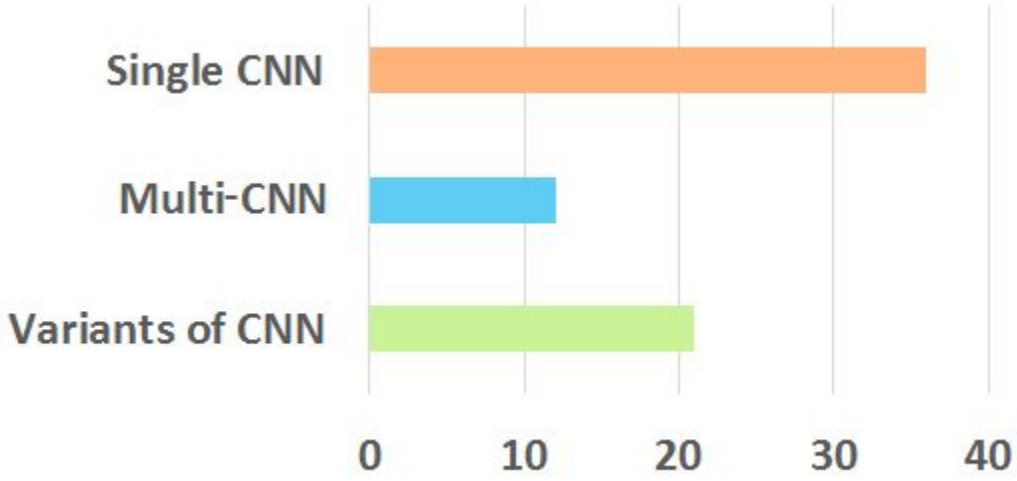
- Methods based on a single deep CNN

❑ Multi-CNN

- Use >= one CNN to extract different deep features and concatenate them as the final face representation

❑ Variants of CNN

- Unconventional CNN based framework



❑ Single CNN

- Strategies used to improve FR performance:

(1) Learn more discriminative features directly, e.g.,

- ✔ inserting into the learning process with joint prototype-based and exemplar-based supervision. [Smirnov et al., 2017]
- ✔ Discriminative Covariance oriented Representation Learning (DCRL) framework. [Wang et al., 2017c]

(2) Fuse different types of face features, e.g.,

- ✔ facial attribute feature (FAF) and face recognition features (FRF). [Hu et al., 2017a]

(3) Utilize metric learning algorithms

- ✔ To learn a good distance metric so that the distance between positive face pairs is reduced and that of negative pairs is enlarged as much as possible.
- ✔ E.g., triplet-based metric learning method. [FaceNet, VGGFace]

(4) Design powerful loss functions

- ✔ An effective approach
- ✔ Quantities of loss functions are proposed
- ✔ Center Loss (Wen et al., 2016b), NormFace (Wang et al., 2017a) , SphereFace (Liu et al., 2017b) - Angular Softmax, CosFace (Wang et al., 2018) - large margin cosine loss, ArcFace (Deng et al., 2018) -Additive Angular Margin Loss, COCO Loss (Liu et al., 2017d), etc.

(5) Adopt proper activation functions

- ✔ Choosing a proper activation function is also important.
- ✔ Sigmoid, Tanh, Rectified Linear Units (ReLU), Leaky Rectified Linear Units (LReLU) (Maas et al., 2013), Parametric Rectified Linear Units (PReLU) (He et al., 2015a), etc.
- ✔ Max-Feature-Map (MFM) (Wu et al., 2015)

- **(6) Others**
  - ✔ contrastive convolution, specifically focuses on the contrastive characteristics between two faces. [Han et al., 2018]
  - ✔ sparse neural connections, can get a good initialization and avoid bad local minima. [Sparse ConvNets, (Sun et al.,2016)]
  - ✔ predictable hash code algorithm, map face samples to Hamming space to enhance the predictability of binary codes. [He et al., 2015b]
  - ✔ discriminative face depth estimation approach , based on above. [Cui et al., 2018a]
  - ✔ capture unique and discriminative pairwise relations among different identities by obtaining local appearance patches around landmark points on the feature map. [PRN, (Kang et al., 2018)]

❑ Multi-CNN

- **Two types:**
- **(1) Extract features in different regions of faces**
  - ✔ Trained on different parts and scales of a face [SIAMESE (Wang et al., 2014)]
  - ✔ Four face regions are cropped for feature extraction [MFRS (Zhou et al., 2015)]
  - ✔ Extract overlapped image patches centered at different landmarks on face region [Baidu (Liu et al., 2015)]
  - ✔ **DeepID series methods** (DeepID, DeepID2, DeepID2+, DeepID3) extract robust features of different local face patches.

- **(2) Extract features of different aspects of faces**
  - ✔ Each CNN takes a pair of whole faces or facial components (forehead, eye, nose and mouth) as input. [FR+FCN (Zhu et al., 2014b)]
  - ✔ Stack Multi-scale Convolution Layer Blocks (MCLBs) to present multi-scale abstraction. [Kang et al. (2017)]
  - ✔ Explore the complementarity of two distinct DCNNs by training them with two different large datasets. [Xiong et al. (2017)]
  - ✔ Exploit the complementary information presented in features generated by different DCNNs for template-based face recognition. [Bodla et al. (2017)]
  - ✔ Concatenated different features of two DCNNs. Each type of feature is a combination of multi-scale representations through the use of auxiliary classifiers. [Lu et al. (2017c)]

❑ Variants of CNN

- Unconventional CNN based framework by
- **(1) Design different layout of multiple CNNs**
  - ✔ Bilinear Convolutional Neural Network (BCNN) [Lin et al., 2015]
  - ✔ Pyramid CNN [Fan et al., 2014]
  - ✔ Guided-CNN [Fu et al., 2017]
  - ✔ Tree-structured CNN [Li et al., 2015a]
- **(2) Modify the way of learning kernels**
  - ✔ PCA to learn filter kernels. [PCANet (Chan et al., 2015) , SPCANet (Tian et al., 2015a), Weighted-PCANet (Huang and Yuan, 2015), MSPCANet (Tian et al., 2016)]
  - ✔ Eigenvectors as filter kernels. [SRDANet (Tian et al., 2015b)]
  - ✔ Local Binary Pattern (LBP) as kernels. [LBPNet) (Xi et al., 2016)]
  - ✔ Kernels are dynamically determined by the spatial distribution of facial landmarks. [Li et al., 2015a]
  - ✔ Kernels conditioned on the present intermediate representation and the activation status in lower layers. [Conditional CNN (Xiong et al., 2015)]

- **(3) Fuse CNNs with other types of modules**
  - ✔ Weighted Nearest Neighbor Classifier. [WNNC (Sim´on et al., 2016)]
  - ✔ Neural Aggregation Network (NAN). [NAN (Yang et al., 2017a)]
  - ✔ Attention-Based Template Adaptation Module. [ABTA (Dong et al., 2017a)]
  - ✔ Recursive Spatial Transformer module. [ReST (Wu et al., 2017a)]
- **(4) Adopt weakly-supervised or unsupervised learning**
  - ✔ Weakly-supervised self-learning DCNN. [SLDCNN (Chen and Deng, 2016)]
  - ✔ unsupervised learning. [LBPNet (Xi et al., 2016) , JFL (Lu et al., 2015a)]
- **(5) Others**
  - ✔ Deal with false positives through employing model uncertainty. [Bayesian DCNN (B-DCNN) (Zafar et al., 2019)]

❖ AE and its Variants based deep methods

- Variants
  - DAE: Denoising Autoencoder
    - Enhances its generalization by training with locally corrupted inputs
    - Does two things:
      - ✔ encode the input
      - ✔ undo the effect of a corruption process
  - SAE: Stacked Autoencoder
    - Stacked to form a deep network by feeding the latent representation of an AE as input to the next AE
  - CAE: Contractive Autoencoder
  - VAE: Variational Autoencoder
- Usage
  - ❑ Learn common latent features between different domains
  - ❑ Reduce the dimension of learned features
  - ❑ Reconstruct images

❑ Learn common latent features for cross-domain FR

- **Universally used in HFR, e.g., cross-age, -large pose, -various expressions.**
  - ✔ CpAEs (Riggan et al., 2015)
  - ✔ Coupled Autoencoder Networks (CAN) (Xu et al., 2017a)
  - ✔ Deep Discriminant Analysis (DDA) Nets (Pathirage et al.,2016)
  - ✔ Random Faces Guided Sparse Many-to-one Encoder (RF-SME) (Zhang et al., 2013)
  - ✔ Stacked Progressive Autoencoder (SPAE) (Kan et al., 2014)
  - ✔ Stacked Face Denoising Autoencoders (SFDAE) (Pathirage et al., 2015)
  - ✔ Distilling and Dispelling Autoencoder (*D*$_2$*AE*) (Liu et al., 2018e)

❖ GAN based deep methods

❑ Face Synthesis

- ✔ DA-GAN (Zhao et al., 2017)
- ✔ Age-cGAN (Antipov et al, 2017b); AgecGAN+LMA (Antipov et al, 2017a)
- ✔ GAN based Visible Face Synthesis (GAN-VFS)

❑ Domain-invariant feature learning

- Video: DAN (Rao et al, 2017a)
- Pose: DR-GAN (Tran et al, 2017) , UV-GAN (Deng et al., 2018a)
- Makeup: BLAN (Li et al., 2018b)
- NIR-VIS: Song et al. (2018)
- HFR: Song et al. (2017) ,Cao et al. (2018a)
- …

❖ Hybrid Architectures

- Combine two or more types of neural networks
  - AE+DBM:
    - ✔ Nagpal et al. (2015)
    - ✔ Goswami et al. (2017)
    - ✔ MDLFace (Goswami et al., 2014)
  - CNN+RBM:
    - ✔ Sun et al. (2013)
    - ✔ McDFR (Chen et al., 2015c)
  - GAN+CNN
    - ✔ Zhang et al. (2017b)

❖ Loss functions

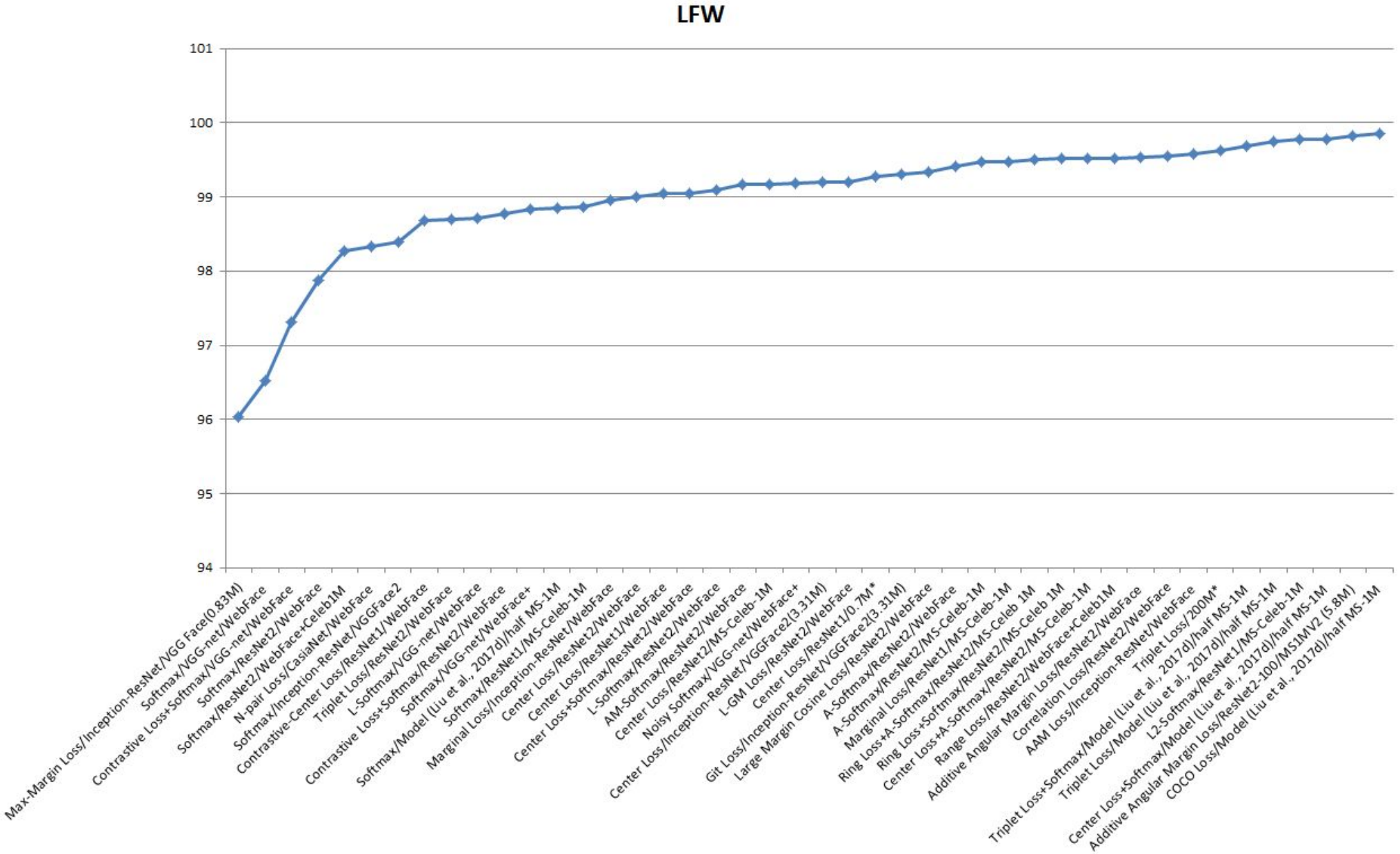
❑ Sample-based supervision

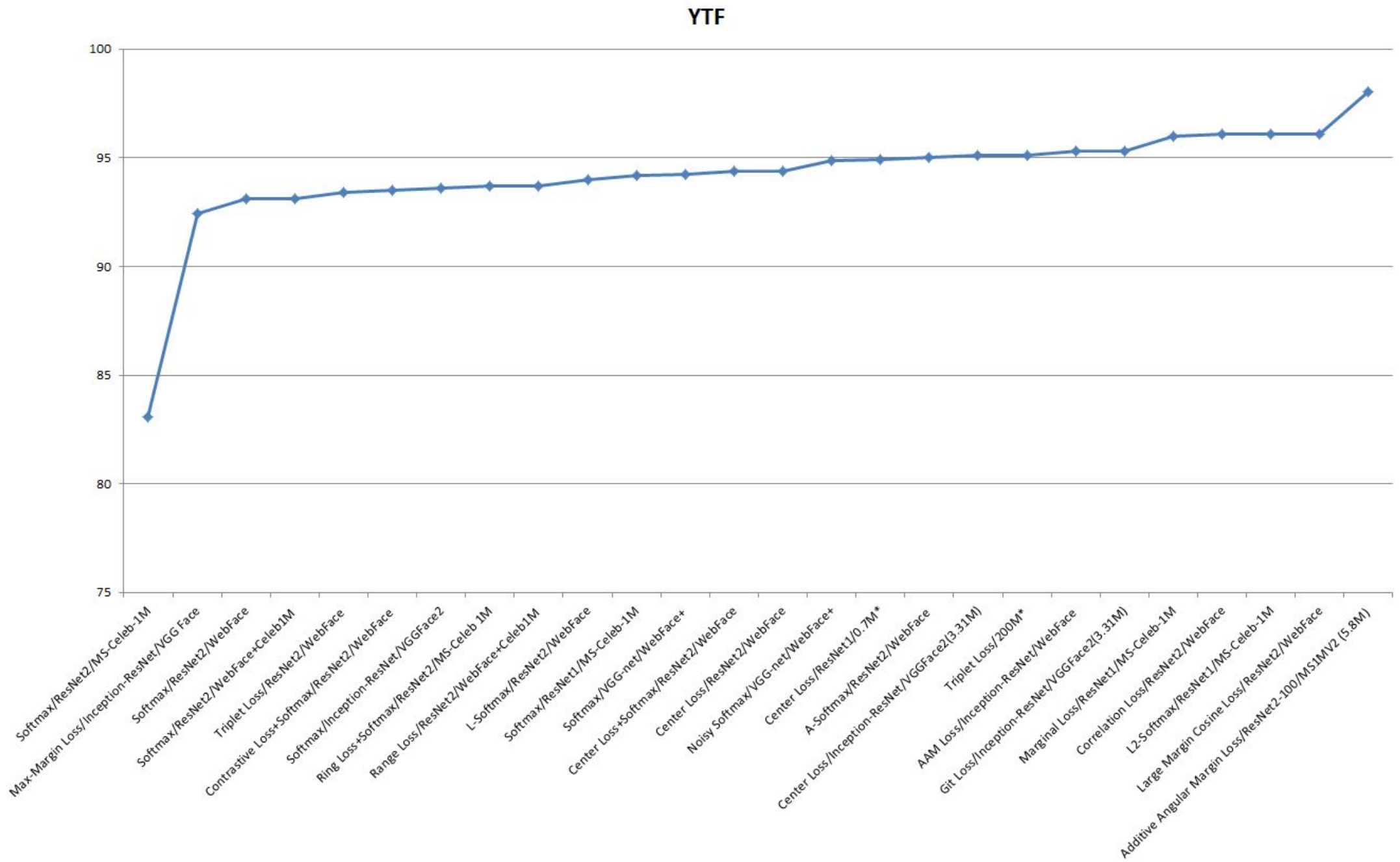
- Softmax, L2-Softmax, triplet loss, marginal loss, COCO loss, Ring loss, Large-margin softmax loss, Angular softmax loss, Adaptive Angular Margin Loss, Additive margin softmax loss, additive angular margin loss, large margin cosine loss

❑ Set-based supervision

- Center loss,contrastive-Center loss, Range loss, Git Loss, Max-margin loss

Comparison of different loss




YTF
100
95
90
85
80
75
Softmax/ResNet2/MS-Celeb-1M
Max-Margin Loss/Inception-ResNet/VGG Face
Softmax/ResNet2/WebFace
Softmax/ResNet2/WebFace+Celeb1M
Triplet Loss/ResNet2/WebFace
Contrastive Loss+Softmax/ResNet2/WebFace
Softmax/Inception-ResNet/VGGFace2
Ring Loss+Softmax/ResNet2/MS-Celeb 1M
Range Loss/ResNet2/WebFace+Celeb1M
L-Softmax/ResNet2/WebFace
Softmax/ResNet1/MS-Celeb-1M
Softmax/VGG-net/WebFace+
Center Loss+Softmax/ResNet2/WebFace
Center Loss/ResNet2/WebFace
Noisy Softmax/VGG-net/WebFace+
Center Loss/ResNet1/0.7M*
A-Softmax/ResNet2/WebFace
Center Loss/Inception-ResNet/VGGFace2(3.31M)
Triplet Loss/200M*
AAM Loss/Inception-ResNet/WebFace
Git Loss/Inception-ResNet/VGGFace2(3.31M)
Marginal Loss/ResNet1/MS-Celeb-1M
Correlation Loss/ResNet2/WebFace
L2-Softmax/ResNet1/MS-Celeb-1M
Large Margin Cosine Loss/ResNet2/WebFace
Additive Angular Margin Loss/ResNet2-100/MS1MV2 (5.8M)

- Overview

- Deep Learning Methods on FR

- Specific Face Recognition Problems
- Databases
- Discussions and Challenges

- In early time, FR research mainly concentrated on:
  - visible light face images
  - video faces

Visible images

Video

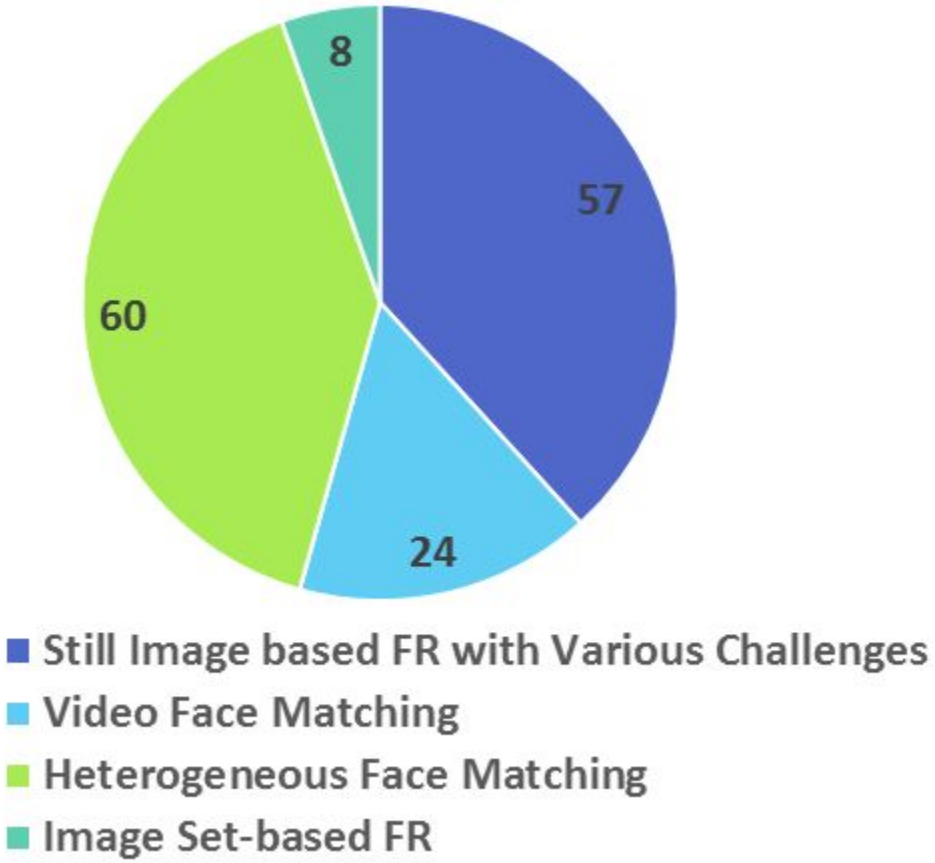
- With the emergence of various types of face data, research concentrations have also focused on some specific tasks:
  - robust to changes of
    - ✔ pose, illumination, occlusion, makeup, expression, cross-age, etc.
  - improving performance of video, heterogeneous FR
    - ✔ e.g., Still-to-Video, NIR-VIS, Sketch-Photo, Cross-Resolution, 3D based, ID-Selfie.

RGB-D

Sketch-photo

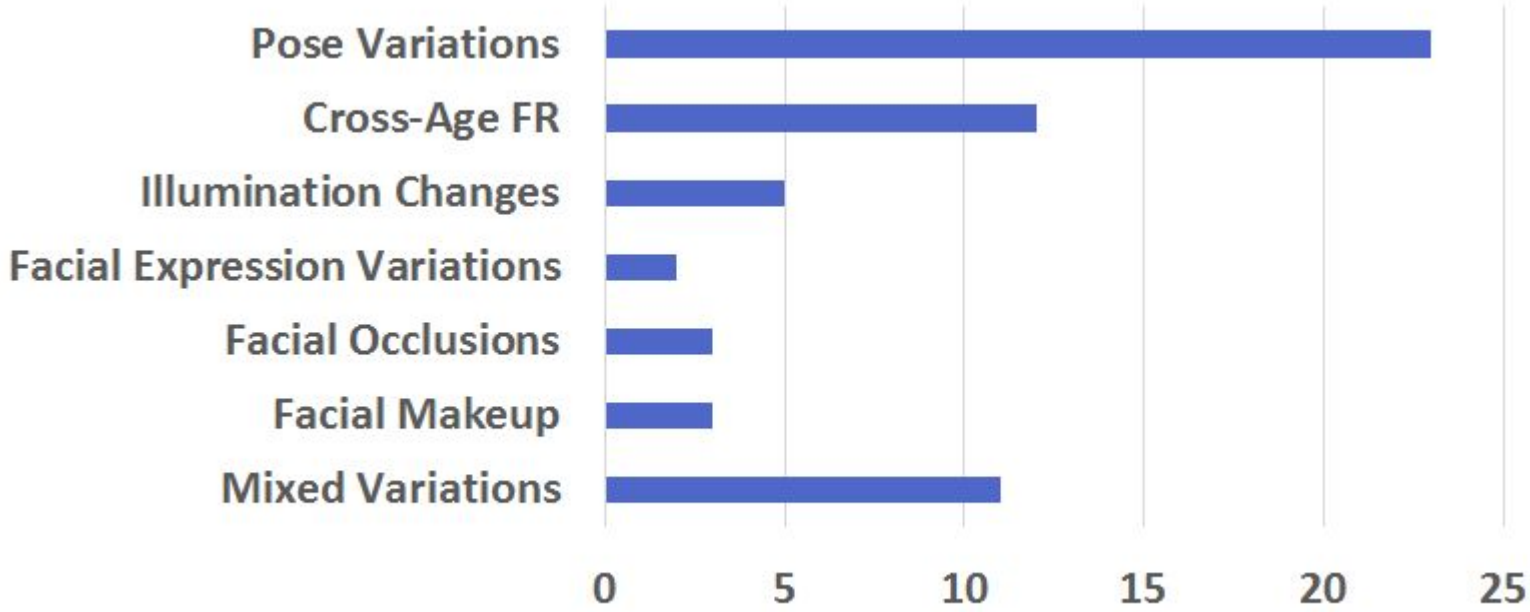
NIR-VIS

- In addition to general FR, there are some FR problems that researchers address specifically with deep learning methods
- We discuss these problems:
  - Some challenges in still image based FR
    - ✔ pose variations, cross-age, illumination changes, etc.
  - Video FR
  - Heterogeneous FR
    - ✔ Still-to-Video
    - ✔ NIR-VIS
    - ✔ Sketch-Photo
    - ✔ Cross-Resolution
    - ✔ 3D based
    - ✔ ID-Selfie
  - Image Set-based FR



❖ Challenges in Still Image-based FR

- In the past decade, face recognition has made significant progress in controlled scenarios, e.g., mugshot
- Recently, researchers focus more on unconstrained face recognition, containing various poses, illuminations, expressions, ages and occlusions



❑ Pose Variations

- Still a challenge for FR, even with deep learning
- Pose-Invariant Face Recognition (PIFR) is far from being solved

- Existing PIFR methods:
  - (1) Employ face frontalization to synthesize a frontal view before feature extraction
    - ✔ Adopt multiple progressive AEs. [Kan et al., 2014]
    - ✔ by learning the displacement field. [Hu et al., 2017b]
    - ✔ Deep Residual EquivAriant Mapping (DREAM) block. [Cao et al., 2018c]
    - ✔ Face Frontalization Generative Adversarial Network (FF-GAN). [Yin et al., 2017]
    - ✔ Two-Pathway Generative Adversarial Network (TP-GAN) [Huang et al. (2017)]

- (2) Directly learn pose-invariant representation from non-frontal face images
  - ✔ Random faces guided sparse many-to-one encoder (RFSME) (Zhang et al., 2013)
  - ✔ 3D-aided 2D face recognition. [Xu et al., 2017b]
  - ✔ Use reconstruction loss to regularize identity feature learning. [Peng et al., 2017]
  - ✔ Make the identity metrics more pose-robust by mitigating the information contained in the pose verification task. [Lu et al., 2017a]
  - ✔ Incorporat a simulator (3D Morphable Model) to obtain shape and appearance prior and leveraged a global local GAN to enhance the realism of both global structures and local details of the face simulators output, while preserving the identity information. [Zhao et al., 2018c]

- (3) Perform both (1) and (2) jointly
  - ✔ perform both tasks jointly to allow them to benefit from each other.
  - ✔ Disentangled Representation learning-GAN (DR-GAN). [Tran et al., 2017]
  - ✔ Pose Invariant Model (PIM). [Zhao et al., 2018b]
  - ✔ Pose-Aware Models (PAM). [Masi et al., 2019a]
- (4) Others.
  - ✔ Pose-Invariant Similarity Index (PISI) model. [Grm et al., 2016]
  - ✔ UV-GAN. [Deng et al., 2018a]

❑ Cross-Age

- With aging, the facial appearance can change significantly
- Four types of methods:
- (1) directly extract age-invariant features
  - ✔ Deep joint metric learning framework. [Li et al., 2015b]
  - ✔ Latent factor guided CNN. [Wen et al., 2016a]
  - ✔ Age estimation task guided CNN. [Zheng et al., 2017a]
  - ✔ Coupled AE network. [Xu et al., 2017a]
  - ✔ Multi-task deep neural network architecture. [Wang et al., 2017d]
  - ✔ Distance metric optimization driven learning approach. [Li et al., 2018c]
  - ✔ Age-related factor guided joint task modeling CNNs. [Li et al., 2018a]

- (2) Synthesize a face that matches target age before feature extraction.
  - ✔ Age-cGAN aging/rejuvenation method. [Antipov et al., 2017b]
  - ✔ Local Manifold Adaptation (LMA) approach. [Antipov et al., 2017a]
- (3) Perform both strategies jointly.
  - ✔ An unified deep architecture jointly learning disentangled identity representations that are invariant to age and performing photorealistic cross-age face image synthesis that can highlight an important latent representation. [Zhao et al., 2018a]

❑ Illumination Changes

- Lighting condition is one of the big factors for facial appearance change and recognition performance degradation
- Illumination changes may cause huge differences of facial shading or shadow from varying directions or energy distributions of the ambient lighting, together with the 3D structure of faces
- Methods:
  - ✔ fuzzy-neural network. [Thakare and Thakare, 2011]
  - ✔ DCNN [Choi et al., 2016]

❑ Partial Face Images

- Partial face images occur when a face is:
  - Occluded by objects, e.g., faces of other individuals, sunglasses, hats, beard, masks or scarves;
  - Captured in various poses without user awareness;
  - Positioned partially out of the camera's filed of view.
- Methods:
  - ✔ Multi-Scale Region-based Convolutional Neural Network (MR-CNN)model. [He et al., 2016b]
  - ✔ Dynamic Feature Matching (DFM) method. [He et al., 2018a]

❑ Facial Makeup

- Makeup brings about remarkable facial appearance changes resulting in both global and local appearance discrepancies between makeup and non-makeup face images.
- Methods:
  - ✔ Most rely much on the various cues and information captured by the effective appearance features, which lack robustness over the application of makeup that is non-permanent as well as miscellaneous. [Guo et al., 2014; Zheng and Guo, 2016]
  - ✔ Bi-level adversarial network (BLAN) , simultaneously considers makeup removal and face verification. [Li et al., 2018b]

❑ Facial Expression Variations

- Facial expression changes also impose problems for FR
- Facial deformations with expressions can change the appearance
- Methods:
  - ✔ Stacked denoising autoencoder for expression-robust feature acquisition. It exploits contributions of different color components in different local face regions by recovering the neutral expression from various other expressions, and processes the faces with dynamic expressions progressively. [Pathirage et al., 2015]
  - ✔ fused 2D images of a face and motion history images (MHIs), which are generated from the same subject's image sequences with expressions to do face recognition. [Liu et al., 2016a]

❑ Mixed Variations

- Deep learning methods are good at:
  - ✔ dealing with nonlinear characteristics in face images
  - ✔ and making the extracted features more discriminative
- Methods proposed to address more than one challenges
- (1) Pose+Illumination.
  - ✔ By rotating a face with any pose and illumination to a canonical view. [FIP (Zhu et al., 2013), MVP (Zhu et al.,2014a) , CPF (Yim et al., 2015)]

- (2) Pose+Expression.
  - ✔ learn dynamic data adaptive features. [Deep Discriminant Analysis (DDA) Nets, (Pathirage et al., 2016)]
  - ✔ disentangle irrelevant non-rigid appearance variations. [Tree-structure Kernel Adaptive CNN, (Li et al., 2015a)]
- (3) Pose+Illumination+Expression.
  - ✔ Yin and Liu (2018)
  - ✔ Ding and Tao (2015).
- (4) Multiple Challenges
  - ✔ Sun et al. (2014a); Zhu et al. (2014b); Hu et al. (2017b)

❖ Video-based Face Recognition (VFR)

- VFR has emerged as an important topic
  - Due to the increasing number of CCTV cameras installed
  - and the easy availability of video recordings
- The image quality of video frames tends to be significantly lower
- Faces exhibit much richer variations, e.g., motion blur, out-of-focus blur, a large range of pose variations
- Surveillance and mobile cameras are often low-cost (and therefore low-quality) devices, which further exacerbates problems with video frames.

- Two types of Methods:

❑ Methods that performed on images can be used on videos

✔ e.g., DDML (Hu et al.,2014), DeepFace (Taigman et al., 2014), DeepID2+ (Sun et al.,2015b), FaceNet (Schroff et al., 2015), Light CNN (Wu et al.,2015), VGGFace (Parkhi et al., 2015), He et al. (2015b), etc.

❑ Methods that specially targeted for VFR

✔ extract discriminative embeddings of still ROI and then compare with ROIs in videos. [Parchami et al., 2017a]

✔ image to video feature-level domain adaptation approach [Sohn et al., 2017]

✔ Attention-Set based Metric Learning method, measure the statistical characteristics of image sets for VFR. [ASML (Hu et al., 2017c)]

✔ Trunk-Branch Ensemble CNN model (TBE-CNN). [Ding and Tao, 2018]

✔ dependency-aware attention control (DAC). [Liu et al. (2018c)]

✔ Kim et al. (2018)

❖ Heterogeneous Face Recognition

- The main challenges lie in the large modality discrepancy, e.g.,
  - comparing single vs. multi-channel imagery
  - linear and non-linear variations in intensity value due to:
    - ✔ different specular reflection properties
    - ✔ different coordinate systems
    - ✔ reduction of appearance detail
    - ✔ non-rigid distortion preventing alignment, etc.
  - insufficient training samples

- Generative HFR methods used for multiple scenarios of face matching between different modalities.

❑ Feature descriptor based methods

- ✔ directly extract modality invariant features for recognition
- ✔ Ding and Tao, 2015; Yi et al., 2015; Saxena and Verbeek, 2016; Kan et al., 2016; Song et al., 2017; Cao et al., 2018b; Peng et al., 2019; Deng et al., 2019.

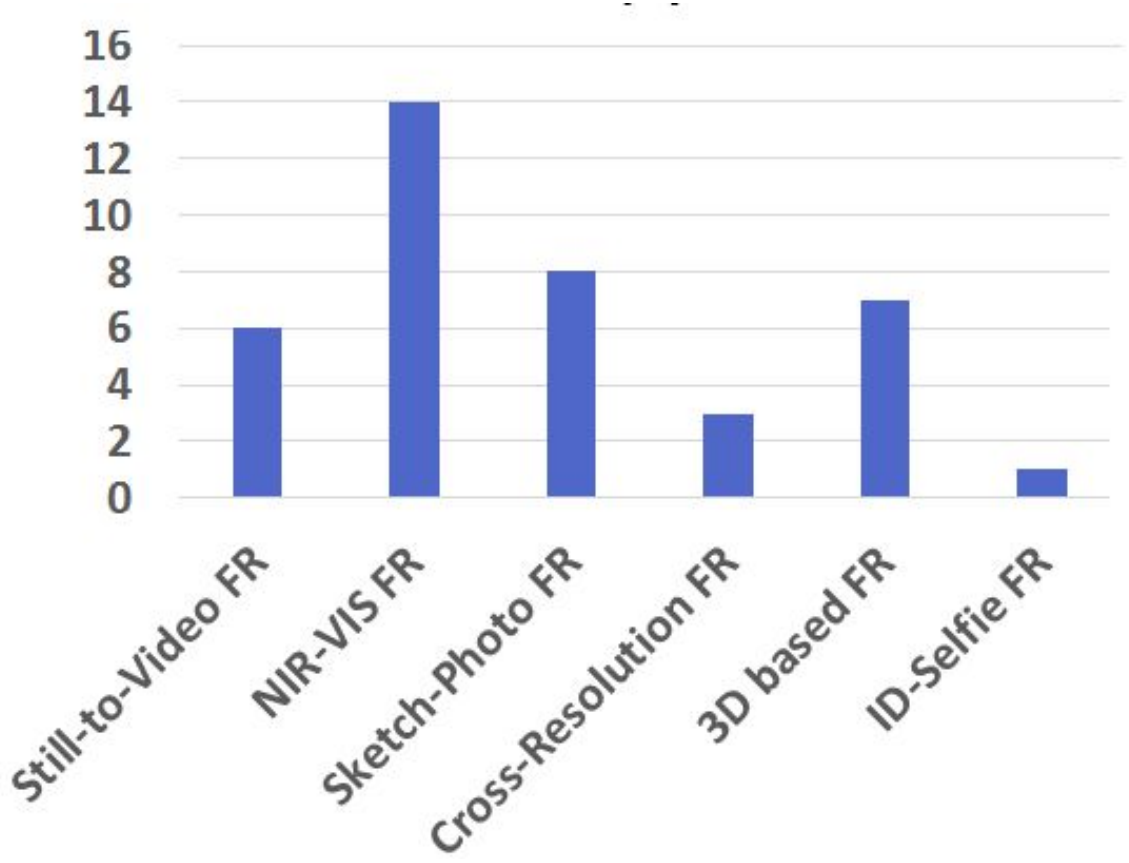
❑ Synthesis based methods

- ✔ firstly transform images in one modality to another, and do conventional homogeneous FR
- ✔ Riggan et al., 2015; Zhang et al., 2017b; Cao et al., 2018a.

❑ Common space projection based methods

- ✔ project heterogeneous face images into a latent common space where the probe image and the gallery images could be matched directly
- ✔ Wu et al., 2017b, 2018b; Liu et al., 2018a.

- Still-to-Video FR
- NIR-VIS FR
- Sketch based FR
- Cross-Resolution FR
- 3D based FR
- ID-Selfie FR



❑ Still-to-Video Face Recognition

- Existing still-to-video face recognition methods mainly contains four categories.
- (1) Feature descriptor based methods
  - ✔ Zhu and Guo (2016); Lin et al. (2017a)
- (2) Synthesis based methods
  - ✔ Canonical Face Representation CNN (CFR-CNN). [Parchami et al., 2017b]
- (3) Common space projection based methods
  - ✔ Bao et al. (2017)
- (4) Others
  - ✔ Savchenko and Belova (2017)

❑ NIR-VIS Face Recognition

- **(1) projecting heterogeneous data onto a common latent space for cross-modal matching**
  - ✔ Reale et al. (2016), He et al. (2018b), Wu et al. (2018a)
  - ✔ Coupled Deep Convolutional Neural Network (CpDCNN), Iranmanesh et al. (2018)
- **(2) extracting domain-invariant features from these modalities**
  - ✔ Riggan et al. (2016a); Liu et al. (2016c); He et al. (2017); Lezama et al. (2017); Song et al. (2018)
- **(3) synthesizing visible faces from NIR faces.**
  - ✔ Riggan et al.(2016b); Zhang et al. (2017a)
  - ✔ Deep Perceptual Mapping (DPM), Sarfraz and Stiefelhagen(2017)
  - ✔ Thermal-to-Visible Generative Adversarial Network (TV-GAN), Zhang et al. (2018)
  - ✔ Attribute Preserved Generative Adversarial Network (AP-GAN), Di et al. (2018)
  - ✔ Cross-spectral Face Completion(CFC), He et al. (2019)

❑ Sketch Based Face Recognition (SBFR)

- More challenging than the classical FR from the deep learning point of view.
  - The heterogeneous nature of sketch and photo modalities.
  - The lack of large databases in order to avoid overfitting and local minima.
- Existing approaches primarily focus on closing the semantic gap between the two domains by:
  - ❑ **transfer learning**
    - ✔ Mittal et al. (2015); Galea and Farrugia (2017)
  - ❑ **designing effective similarity measures**
    - ✔ Lin et al. (2017a)
  - ❑ **using facial attributes in conjunction with sketch**

❑ Cross-Resolution Face Recognition

- Low resolution (LR) face images can degrade the face recognition performance significantly
- The simplest solution is to up-scale the probe images, or down-sample the HR images, but it is possible to do better.

- **Synthesis based method**
  - ✔ reconstruct the HR probe image from the LR one by super-resolution (SR) techniques and use it for classification.
- **Projection-based method**
  - ✔ simultaneously transform the LR probe and corresponding HR gallery images into a common feature subspace where the distance between them is minimized.
  - ✔ deep coupled ResNet (DCR) model. [Lu et al. (2018)]

❑ 3D based Face Recognition

- Most deep learning methods are mainly for 2D face recognition
- With the advances of 3D sensors, e.g. the Kinect, and point cloud library (PCL)
  - the information of geometric coordinates of real-world objects can be easily collected
  - more three-dimensional volume data can be processed to mitigate the problem associated with 2D images

pointcloudlibrary

- The RGB-D cameras usually provide synchronized images of both color and depth
  - The color image characterizes the appearance and texture information of a face
  - The depth image provides the distance of each pixel from the camera, representing the face geometry to a certain degree
- 3D information represents more discriminative features by the virtue of increased dimensionality

- Methods:
  - ✔ 3D face recognition model. [Kim et al. (2017) ; Jhuang et al. (2016)]
  - ✔ RGB-D images(color and depth images), to achieve a more accurate recognition. [Thakare and Thakare, 2011; Lee et al., 2016; Simon et al. (2016); Liu et al. (2017a) ; Zulqarnain Gilani and Mian, 2018]
- Although 3D face recognition has advantages over its 2D counterpart, it has not yet been fully benefited from the recent developments in deep learning, due to:
  - the unavailability of large training sets
  - large test datasets.
- Besides, the high cost of specialized 3D sensors limits their use in practical applications.

❑ ID-Selfie Face Recognition

- Identity verification plays an important role in our daily lives
- Numerous activities require to verify who we are by showing our ID documents containing face images, e.g., passports and driver licenses.
  - transactions, access to services, transportation, etc.

✔ DocFace (Shi and Jain, 2019) is a domain-specific network to match scanned or digital ID document photos to digital camera photos of live faces by employing a transfer learning technique.

❖ Image Set-based FR(ISFR)

- Compared with the single image based methods, ISFR deals with severe changes of appearance and makes decisions by comparing the query set with gallery sets
- Users supply a set of images of the same unknown individual
- Methods based on image sets are expected to give a better performance than those based on single images
- Video based recognition can be treated as a special case of image set classification

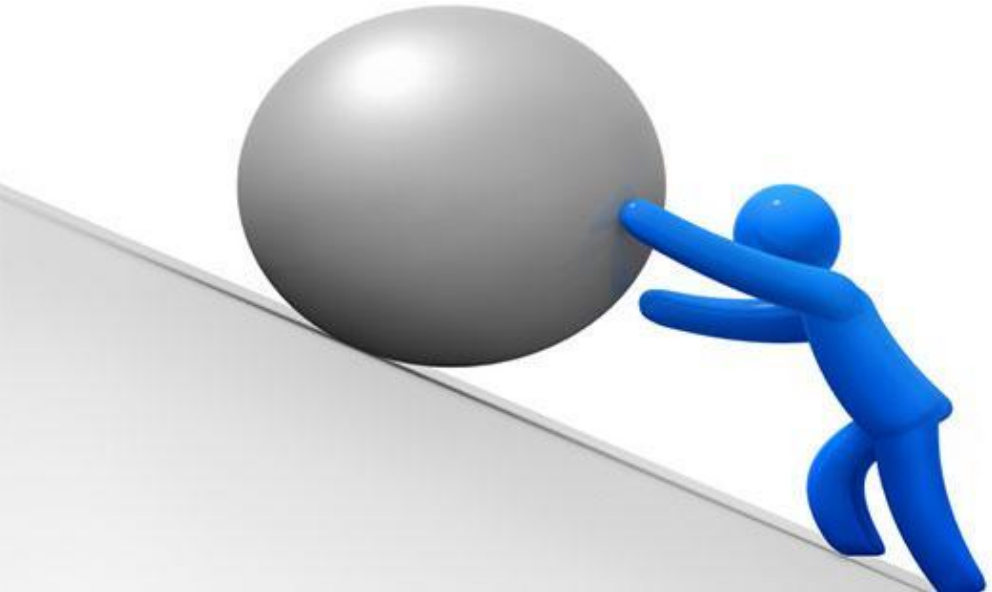
❖ Hard Mining

- Hard mining, previously called bootstrapping, includes hard positive mining and hard-negative mining
- Using “hard” samples can help to improve the decision boundary of the model
- Hard mining is commonly used in object detection
- In FR, some use it to improve training discrimination
  - ✔ Schroff et al., 2015
  - ✔ Parkhi et al., 2015
  - ✔ Zhang et al., 2016b

❖ Closed-Set vs. Open-Set Face Recognition

- Face verification (FV)
  - to determine whether a pair of face images belongs to the same subject
- Face identification (FI)
  - a one-to-many matching
  - usually assuming the query person was already enrolled in the gallery, which is a closed-set problem
- Watch-list
  - similar to face identification
  - but it does not guarantee all query subjects are already registered in gallery, which is an open-set problem

- In the real world, it is normal to treat FI as an open-set problem
- Although FV or closed-set FI has gained good performance, open-set FI is still a challenge



- Overview

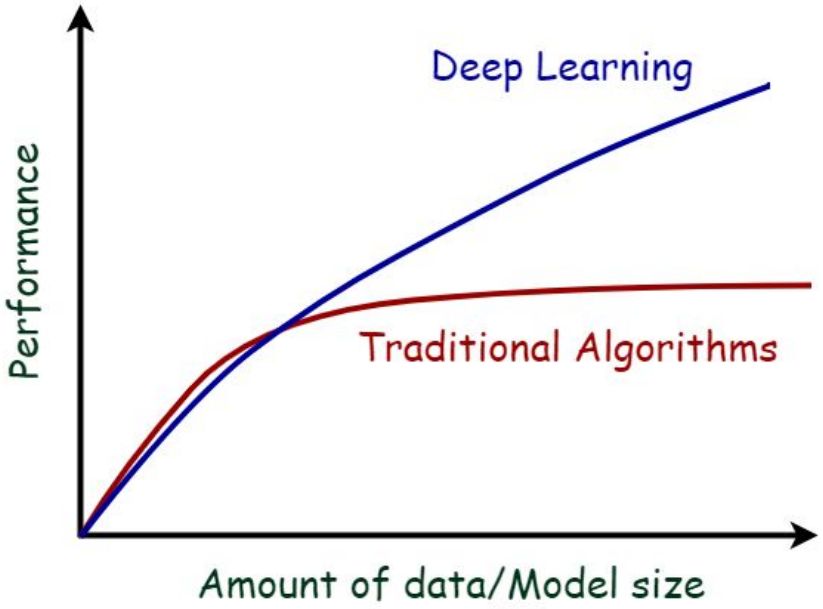
- Deep Learning Methods on FR

- Specific Face Recognition Problems

- Databases

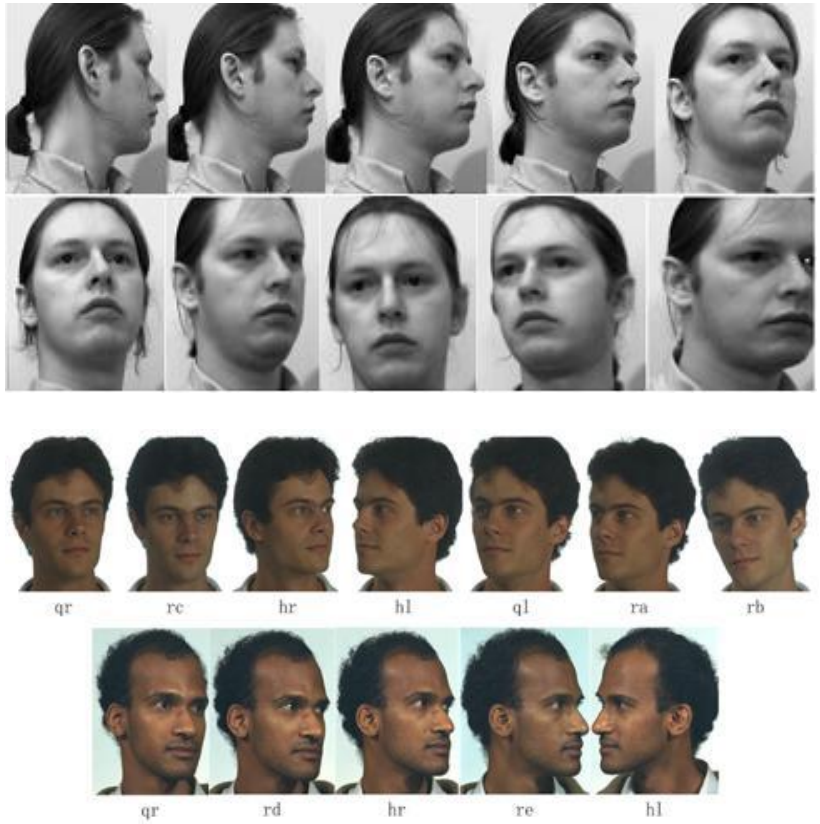
- Discussions and Challenges

- With the wider use of DNNs in FR, the requirement of a huge amount of training data becomes more urgent
- Experiments demonstrated that:
  - large amount of data can help the network learn better models



❖ Still Image Face Databases

- Early datasets were almost collected under pre-defined or controlled environments (PIE, Yale, CMU PIE, FERET, UHDB31, PaSC, COX face, etc.)
- The numbers of subjects/images is small

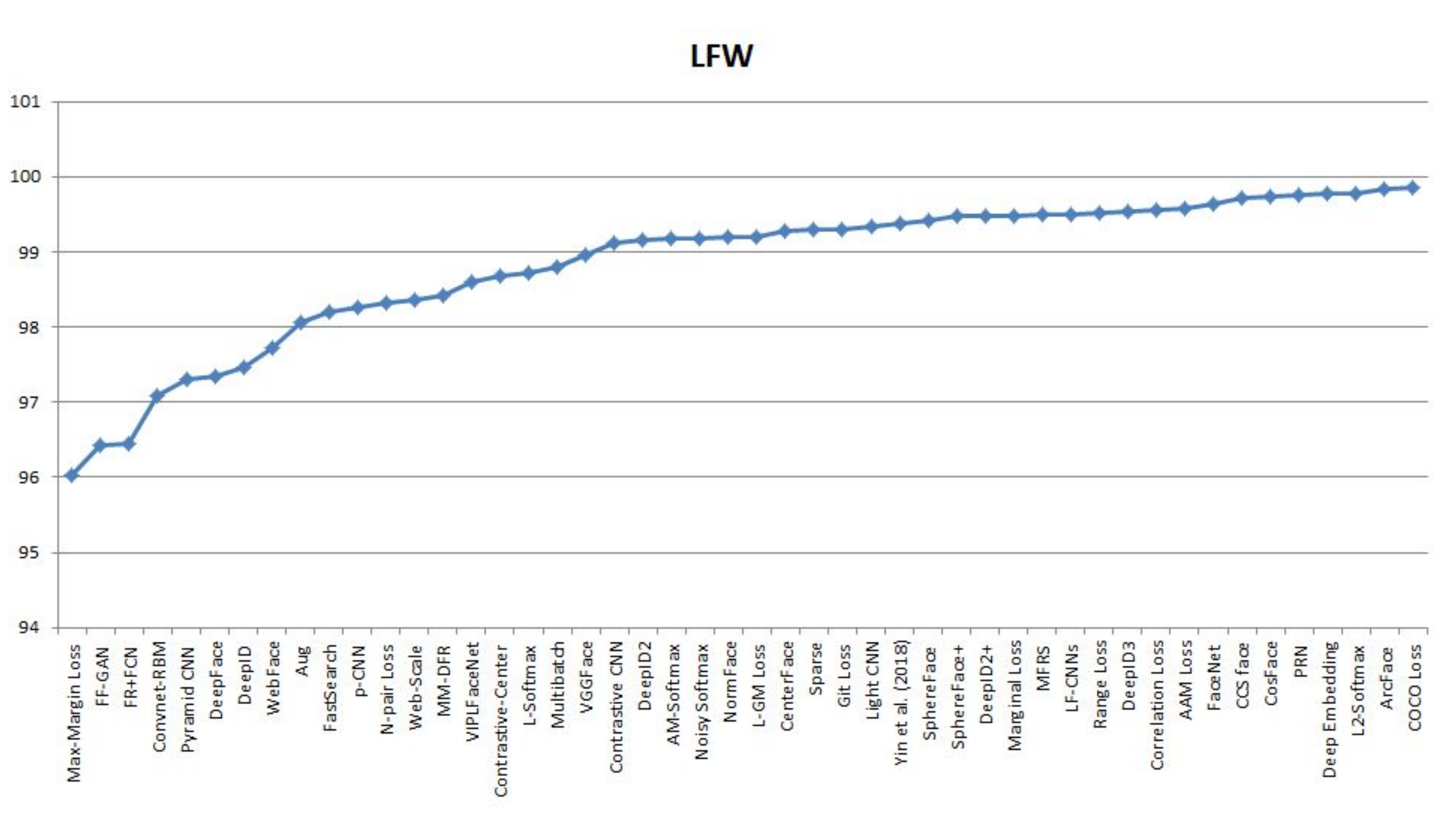


- Along with the practical requirement, more attentions are paid to uncontrolled / unconstrained scenarios (LFW, CACD, IJB-A, CelebFace+, FaceScrub, VGGFace2, etc.)
- #identities/ #images are large

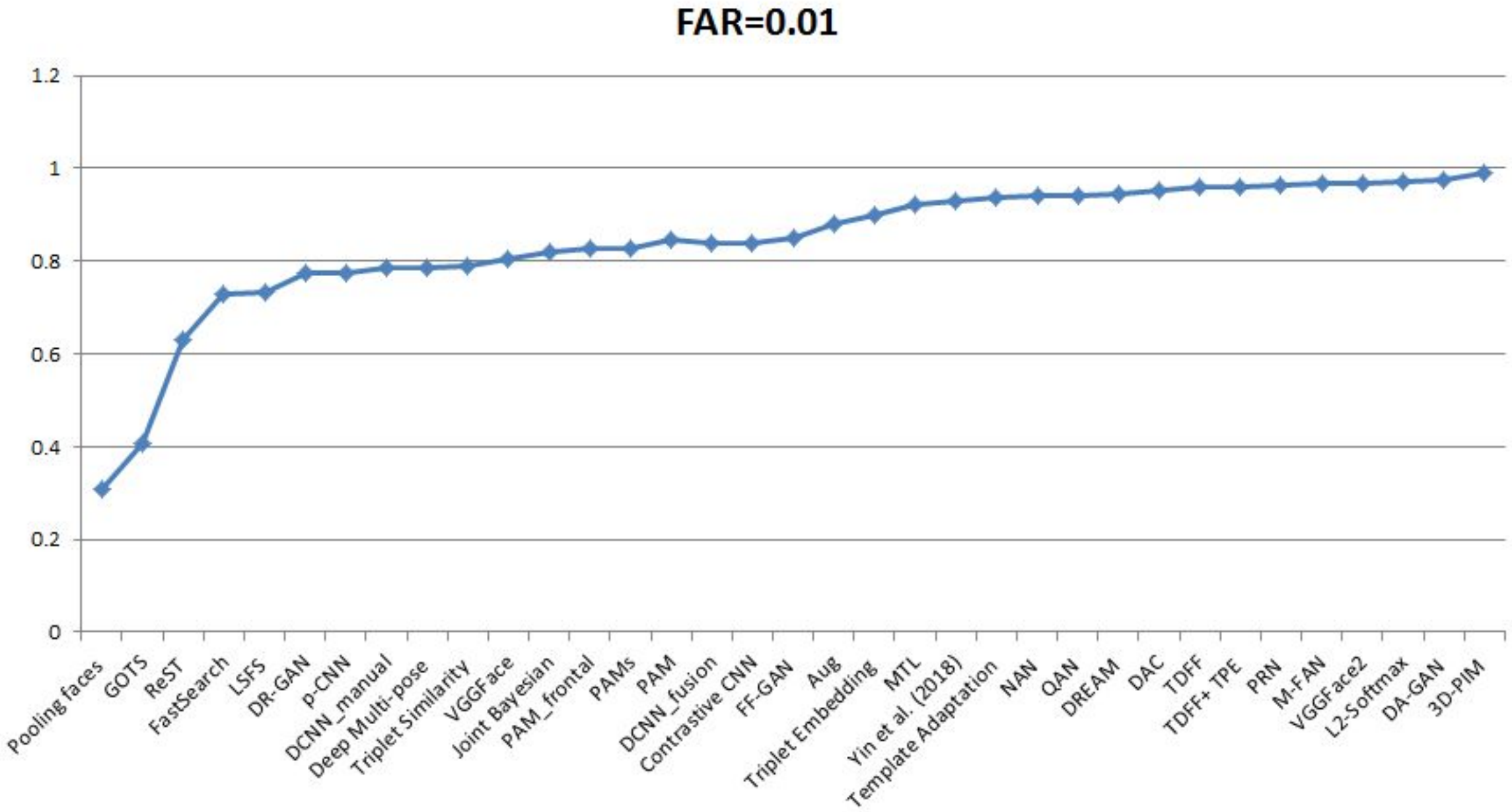
- LFW
  - can be viewed as a milestone dataset in which images are crawled from the Internet containing variations in pose, illumination, expression, resolution, etc.
  - Many research works have been focused on improving the performance on LFW
  - Recent advances, especially the CNN based face recognition, enabled close to 100% accuracy in LFW

However, the face recognition problem is far from being solved

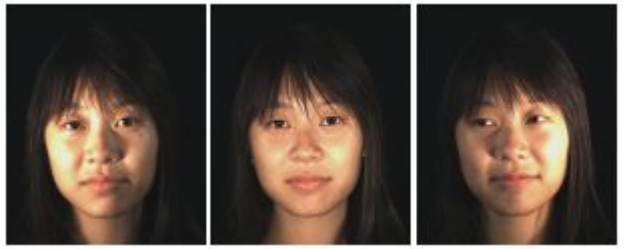
- IJB-A
  - The performance of state-of-the-art face recognition systems are far less than satisfactory on a newly released dataset, IJB-A.
  - This benchmark is considered more challenging than LFW and has become more popular

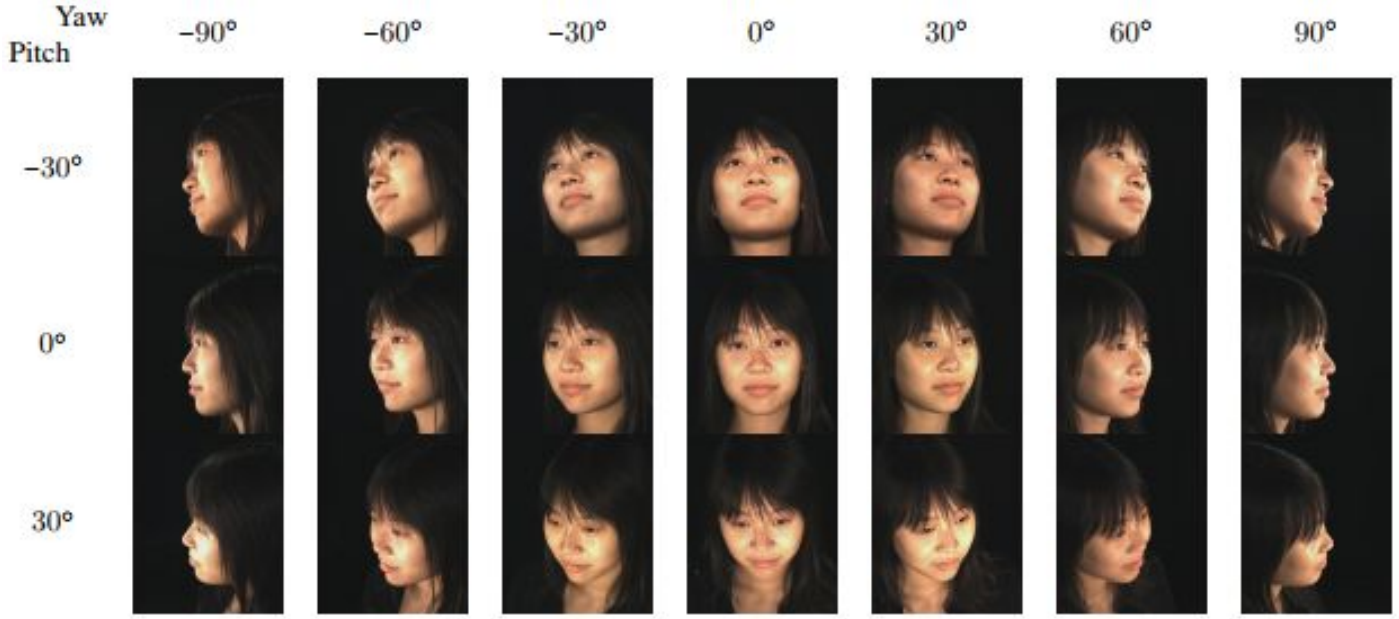

LFW
101
100
99
98
97
96
95
94
Max-Margin Loss
FF-GAN
FR+FCN
Convnet-RBM
Pyramid CNN
DeepFace
DeepID
WebFace
Aug
FastSearch
p-CNN
N-pair Loss
Web-Scale
MM-DFR
VIPLFaceNet
Contrastive-Center
L-Softmax
Multibatch
VGGFace
Contrastive CNN
DeepID2
AM-Softmax
NoisySoftmax
NormFace
L-GM Loss
CenterFace
Sparse
Git Loss
Light CNN
Yin et al. (2018)
SphereFace
SphereFace+
DeepID2+
Marginal Loss
MFRS
LF-CNNs
Range Loss
DeepID3
Correlation Loss
AAM Loss
FaceNet
CCS face
CosFace
PRN
Deep Embedding
L2-Softmax
ArcFace
COCO Loss

- Face verification on IJB-A



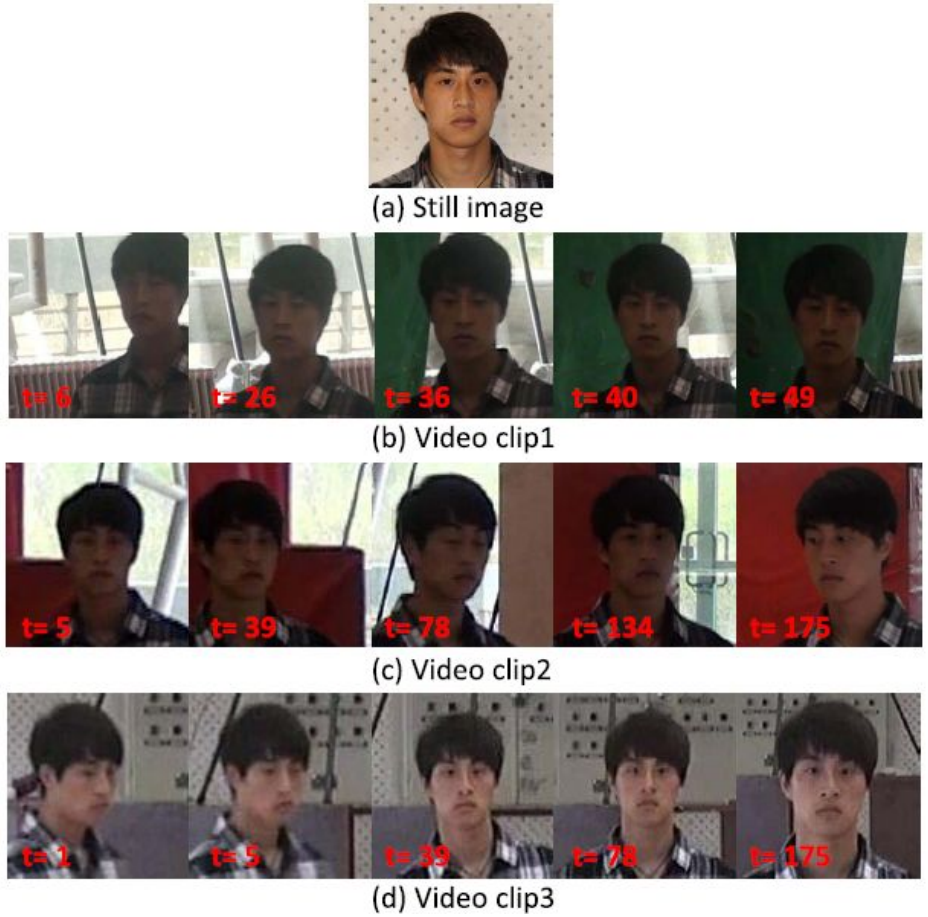
- Some datasets are used for specific tasks, e.g., age, pose, illumination, expression.



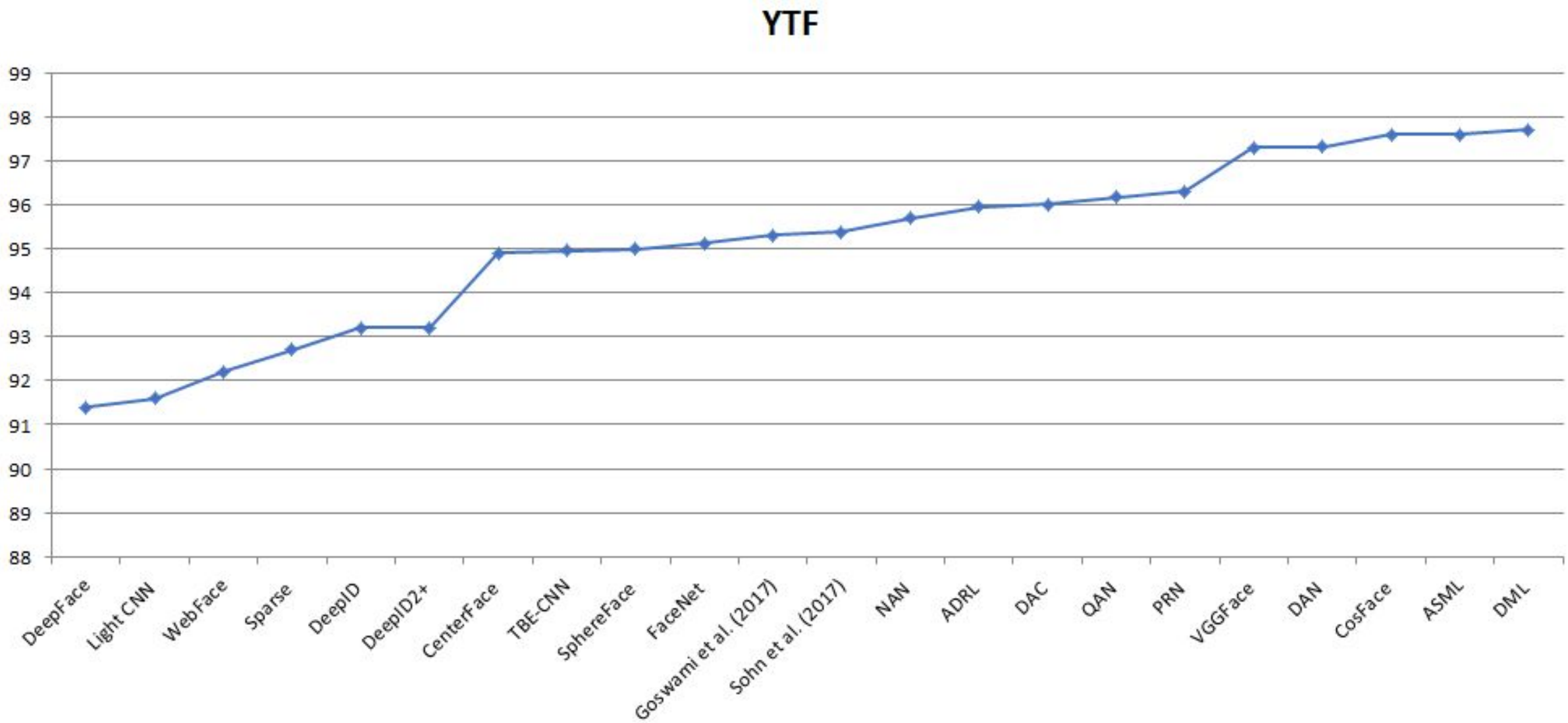


❖ Video Face Databases

- Video based face recognition has also gained much attention
- Most are public available
- Still + Video faces: COX Face, PaSC, JANUS CS2 and IJB-A



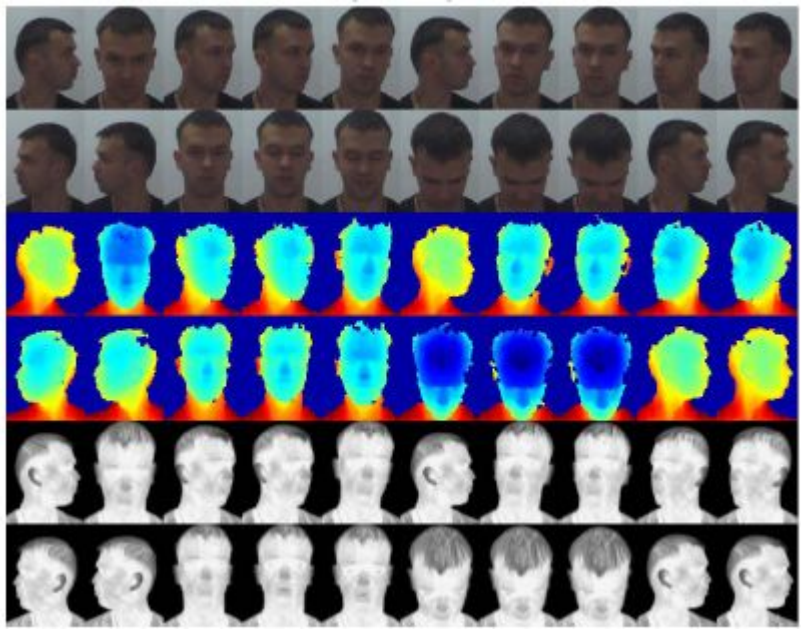
- YouTube Faces (YTF) and YouTube Celebrities (YTC) are often used to test the recognition performance of various deep models.


YTF
99
98
97
96
95
94
93
92
91
90
89
88
DeepFace
Light CNN
WebFace
Sparse
DeepID
DeepID2+
CenterFace
TBE-CNN
SphereFace
FaceNet
Goswami et al. (2017)
Sohn et al. (2017)
NAN
ADRL
DAC
QAN
PRN
VGGFace
DAN
CosFace
ASML
DML

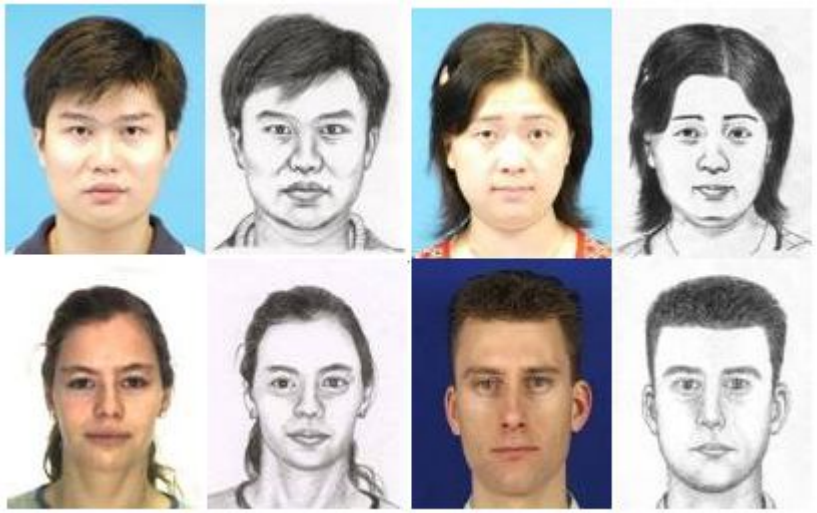
❖ Heterogeneous Face Databases

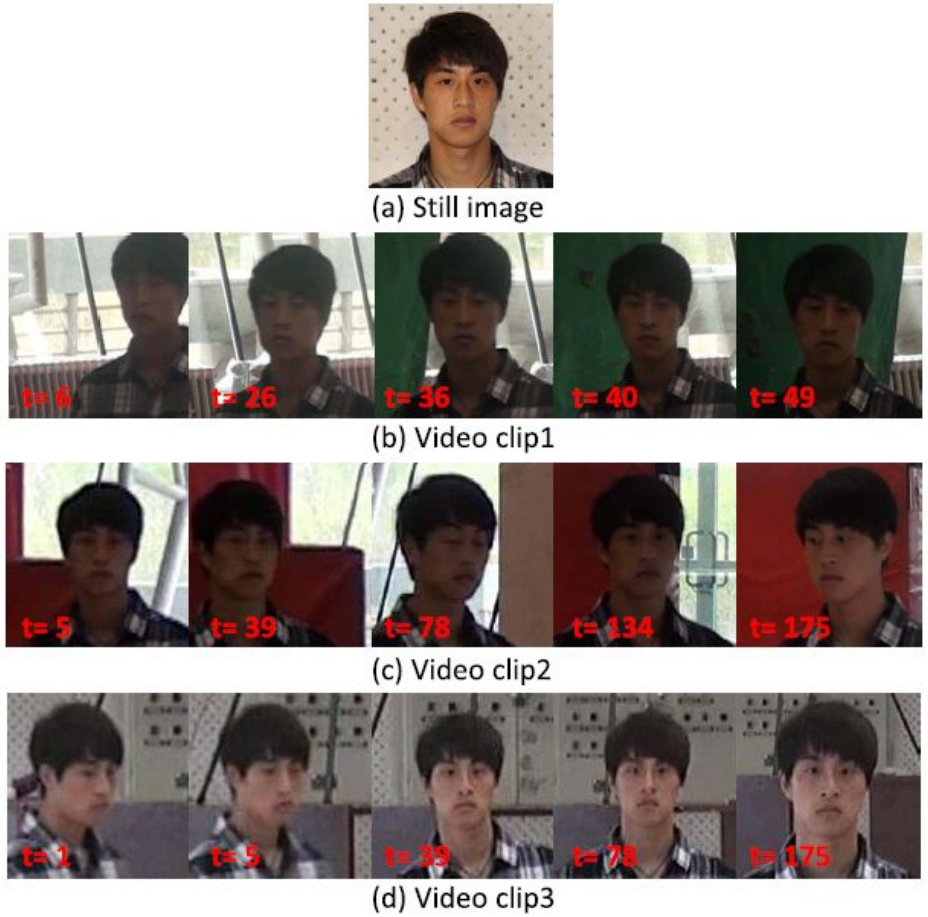
For HFR, multi-modal data are needed

- **Visible and Thermal**
  - ✔ WSRI
  - ✔ Oulu-CASIA
  - ✔ CASIA NIR-VIS 2.0
  - ✔ Night Vision (NVESD)
- **3D**
  - ✔ Curtin-Faces, LS3DFace
  - ✔ RGB-D-T, RGB-D
- **Still and Video**
  - ✔ COX-S2V



- **Sketch-Photo**
  - ✔ CUHK Face Sketch (CUFS)
  - ✔ CUHK Face Sketch FERET (CUFSF)
- **Cross-resolution**
  - ✔ NJU-ID
- **ID-Selfie**
  - ✔ ID-Selfie-A, ID-Selfie-B





•Overview

•Deep Learning Methods on FR

•Specific Face Recognition Problems

•Databases

•Discussions and Challenges

❖ Network Design and Architecture Optimization

- Needs a lot of human interventions in the model design and optimization
- Requires a lot of expert knowledge and takes ample time
- Usually involves a lot of trial and error
- Four methods have been applied to find out hyper-parameters of deep model
  - ✔ Manual Search
  - ✔ Grid Search
  - ✔ Random Search
  - ✔ Bayesian Optimization

❖ Face Data

It is not trivial to get a huge amount of labeled face data

- Some strategies have been developed to address this issue
  - **Minimize the need of data**
    - Peng et al (2016)
    - used a modeling method to minimize the need of huge amount of data
  - **Data synthesis**
    - Lv et al (2017)
    - provided five data augmentation methods for face images, such as landmark perturbation, hairstyles, glasses, poses and illuminations synthesis
  - Instead of directly manipulating the input images, Leng et al (2017) performed virtual sample generation at the feature level for handling unbalanced training set

- For both still and video FR, large-scale datasets are important
- However, large-scale datasets often contain massive noisy labels, especially when automatically collected from the Internet
- Web-collected data could be unbalanced, where some subjects have much more faces than some others

- Unlike 2D images, 3D facial scans are not easy to crawl from the web

- With the progress in sensor technology, low cost 3D sensors may pave the way for multimodal systems, such as color and depth (RGB-D)

Gilani and Mian (2017) proposed:

- ✔A method for generating a large corpus of labeled 3D face identities and their multiple instances for training the models
- ✔A protocol for merging the most challenging existing 3D datasets for testing

- In heterogeneous face recognition, the datasets are typically small
- Developing deep models is likely to overfit or underfit due to the small training set for HFR
- Exploring optimal methods to fit deep models for small-scale HFR datasets remains a critical problem

Thank You and Questions