A Survey on Deep Learning based Face Recognition

Na Zhang

Part III: Specific Problems in Still images, Video and 3D FR

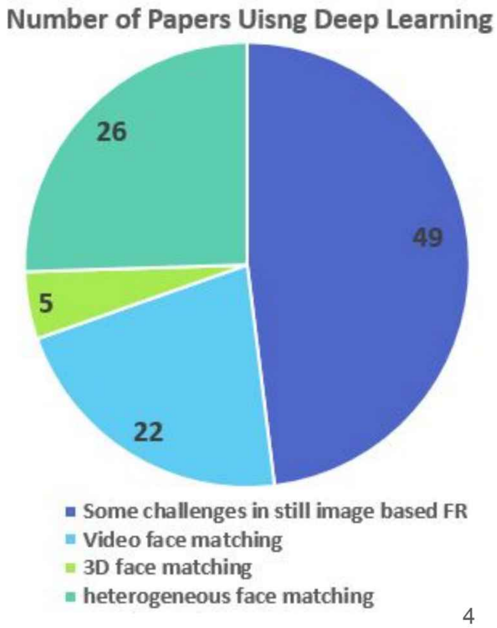
✓ •Introduction

✓ •Deep Learning Methods

•Some Specific Face Recognition Problems

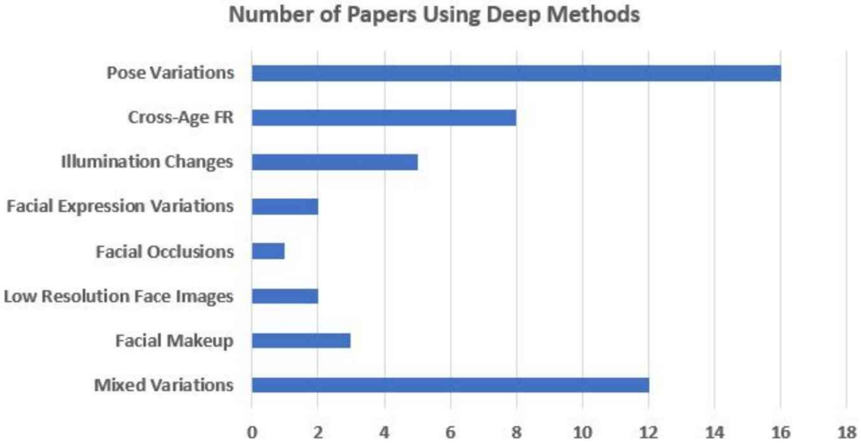
•Databases

- 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
  - □ 3D FR
  - □ Heterogeneous FR
    - ✔ NIR/IR-VIS
    - ✔ photo-sketch
    - ✔ still-to-video



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:
  - employ face frontalization to synthesize a frontal face
  - learn a pose-invariant representation directly from non-frontal face images
  - others

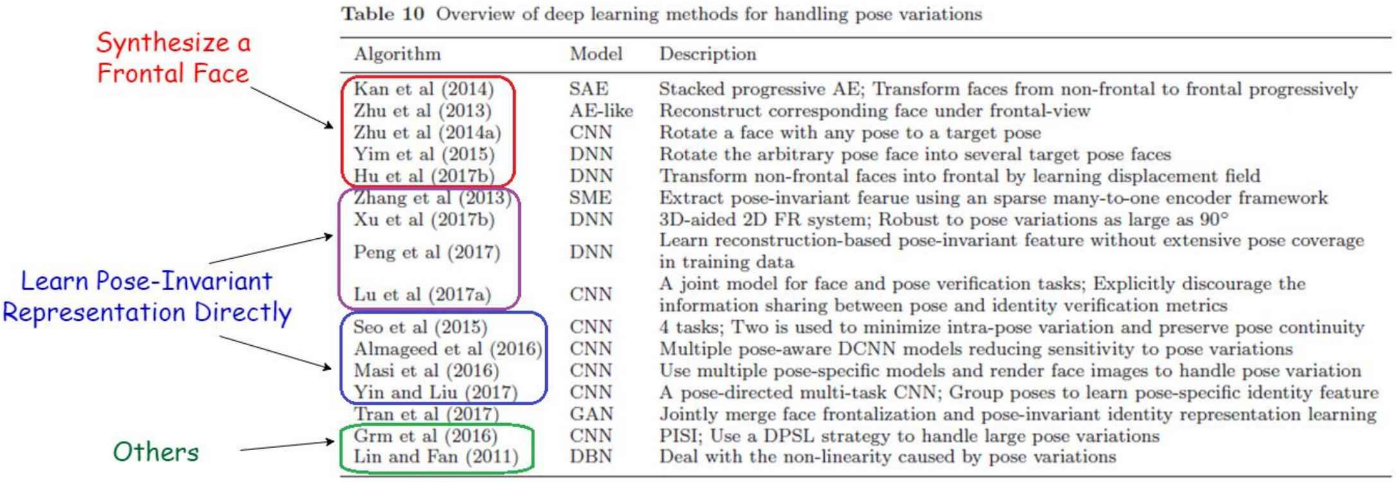

Synthesize a Frontal Face
Learn Pose-Invariant Representation Directly
Others

Table 10 Overview of deep learning methods for handling pose variations

Algorithm	Model	Description
Kan et al (2014)	SAE	Stacked progressive AE; Transform faces from non-frontal to frontal progressively
Zhu et al (2013)	AE-like	Reconstruct corresponding face under frontal-view
Zhu et al (2014a)	CNN	Rotate a face with any pose to a target pose
Yim et al (2015)	DNN	Rotate the arbitrary pose face into several target pose faces
Hu et al (2017b)	DNN	Transform non-frontal faces into frontal by learning displacement field
Zhang et al (2013)	SME	Extract pose-invariant fearue using an sparse many-to-one encoder framework
Xu et al (2017b)	DNN	3D-aided 2D FR system; Robust to pose variations as large as 90°
Peng et al (2017)	DNN	Learn reconstruction-based pose-invariant feature without extensive pose coverage in training data
Lu et al (2017a)	CNN	A joint model for face and pose verification tasks; Explicitly discourage the information sharing between pose and identity verification metrics
Seo et al (2015)	CNN	4 tasks; Two is used to minimize intra-pose variation and preserve pose continuity
Almageed et al (2016)	CNN	Multiple pose-aware DCNN models reducing sensitivity to pose variations
Masi et al (2016)	CNN	Use multiple pose-specific models and render face images to handle pose variation
Yin and Liu (2017)	CNN	A pose-directed multi-task CNN; Group poses to learn pose-specific identity feature
Tran et al (2017)	GAN	Jointly merge face frontalization and pose-invariant identity representation learning
Grm et al (2016)	CNN	PISI; Use a DPSL strategy to handle large pose variations
Lin and Fan (2011)	DBN	Deal with the non-linearity caused by pose variations

- Synthesize a Frontal Face

The ability of generating a realistic frontal face can be beneficial to deal with pose to some extent.

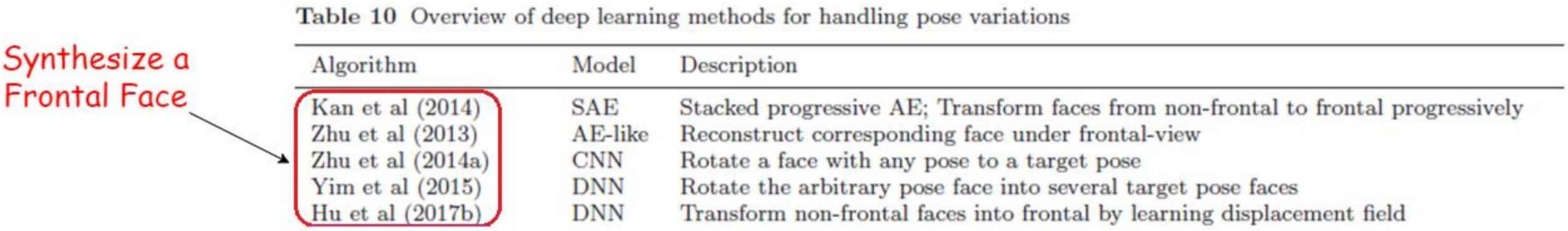

Synthesize a Frontal Face

Table 10 Overview of deep learning methods for handling pose variations

Algorithm	Model	Description
Kan et al (2014)	SAE	Stacked progressive AE; Transform faces from non-frontal to frontal progressively
Zhu et al (2013)	AE-like	Reconstruct corresponding face under frontal-view
Zhu et al (2014a)	CNN	Rotate a face with any pose to a target pose
Yim et al (2015)	DNN	Rotate the arbitrary pose face into several target pose faces
Hu et al (2017b)	DNN	Transform non-frontal faces into frontal by learning displacement field

Kan et al (2014)

used multiple progressive autoencoders to do face frontalization

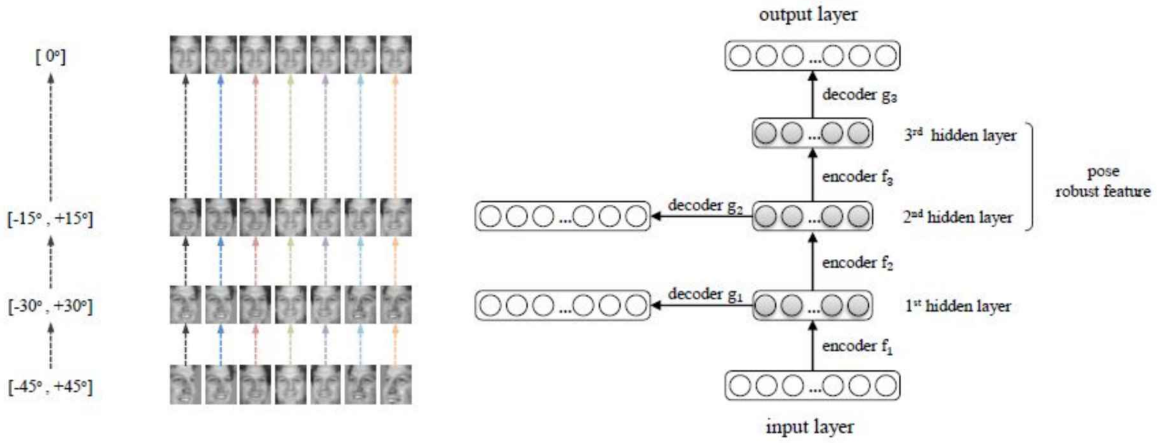


Figure 1. The schema of the proposed Stacked Progressive Auto-Encoders (SPAE) network for pose-robust face recognition. We illustrate an exemplar architecture of the stacked network with $L = 3$ hidden layers, which can deal with poses in yaw rotation within [-45°, +45°]. In training stage of our SPAE, each progressive auto-encoder aims at converting the face images at large poses to a virtual view at a smaller pose (*i.e.*, closer to frontal), and meanwhile keeping the face images with smaller poses unchanged. For instance, for the first progressive AE demonstrated in this figure, only images with yaw rotation larger than 30° are converted to 30°, while other face images with yaw rotation smaller than 30° are mapped to themselves. Such a progressive mode endows each progressive AE a limited goal matching its capacity. In the testing stage, given an image, it is fed into the SPAE network, and the outputs of the topmost hidden layers with very small pose variations are used as the pose-robust features for face recognition.

Kan M, Shan S, Chang H, Chen X (2014) Stacked progressive autoencoders (spae) for face recognition across poses. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 1883–1890

Face Identity-Preserving (FIP) (Zhu et al, 2013)

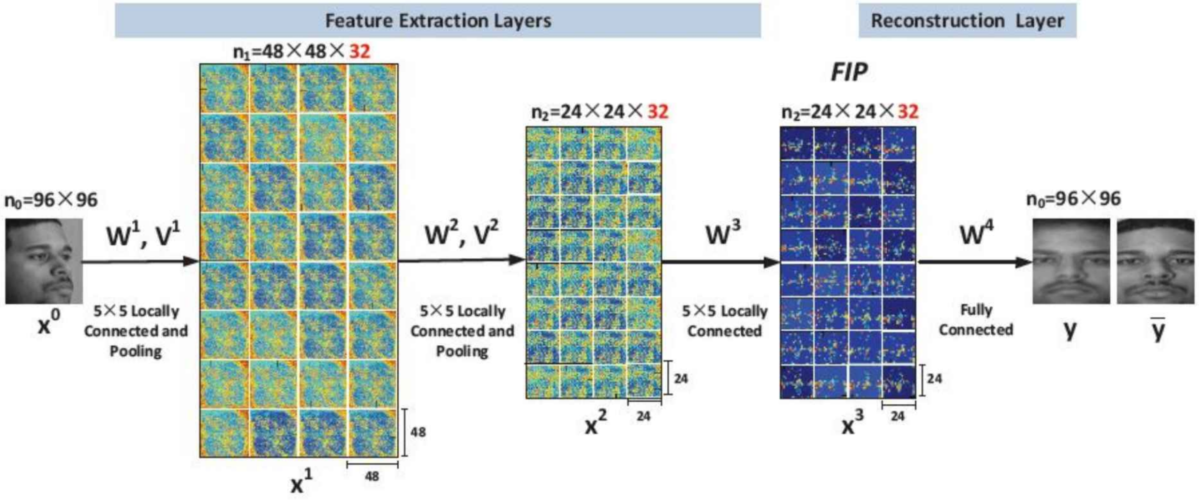


Figure 3. Architecture of the deep network. It combines the feature extraction layers and reconstruction layer. The feature extraction layers include three locally connected layers and two pooling layers. They encode an input face $\mathbf{x}^0$ into FIP features $\mathbf{x}^3$. $\mathbf{x}^1$, $\mathbf{x}^2$ are the output feature maps of the first and second locally connected layers. FIP features can be used to recover the face image $\mathbf{y}$ in the canonical view. $\overline{\mathbf{y}}$ is the ground truth. **Best viewed in color.**

Zhu Z, Luo P, Wang X, Tang X (2013) Deep learning identity-preserving face space. In: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp 113–120

Multi-View Perceptron (MVP) (Zhu et al, 2014a)

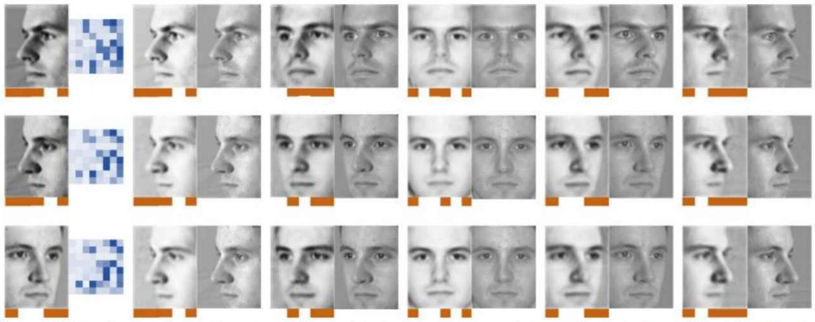


Figure 1: *The inputs (first column) and the multi-view outputs (remaining columns) of two identities. The first input is from one identity and the last two inputs are from the other. Each reconstructed multi-view image (left) has its ground truth (right) for comparison. The extracted identity features of the inputs (the second column), and the view features of both the inputs and outputs are plotted in blue and orange, respectively. The identity features of the same identity are similar, even though the inputs are captured in diverse views, while the view features of the same viewpoint are similar, although they are from different identities. The two persons look similar in the frontal view, but can be better distinguished in other views.*

Zhu Z, Luo P, Wang X, Tang X (2014a) Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems, pp 217–225

Controlled Pose Feature (CPF) (Yim et al, 2015)

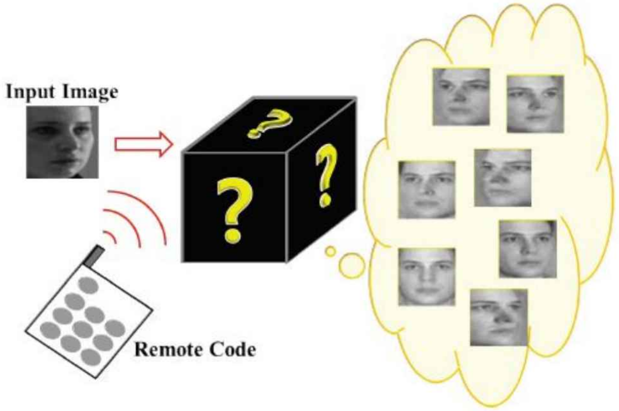
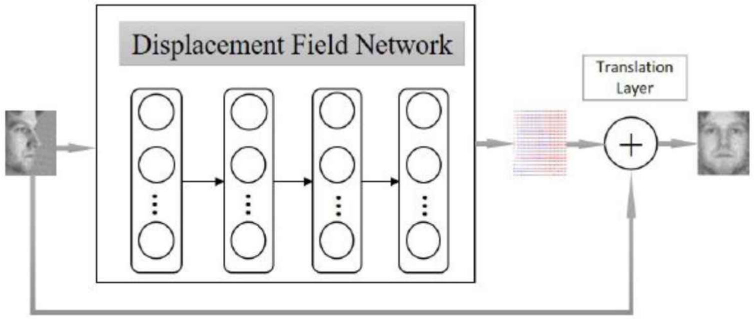


Figure 1. Conceptual diagram of our proposed model. The Input image under an arbitrary pose and illumination is transformed into another pose image. The Remote Code represents the target pose code corresponding to the output image. By interacting between the input image and the Remote Code, our model produces desired pose image.

Yim J, Jung H, Yoo B, Choi C, Park D, Kim J (2015) Rotating your face using multi-task deep neural network. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 676–684

Hu et al (2017b)

- proposed an end-to-end deep neural network
- used to transform a non-frontal face image into a frontal view
- by learning the displacement field,
- which reflects the shifting relationship of pixels from the non-frontal face image and the transformed frontal view

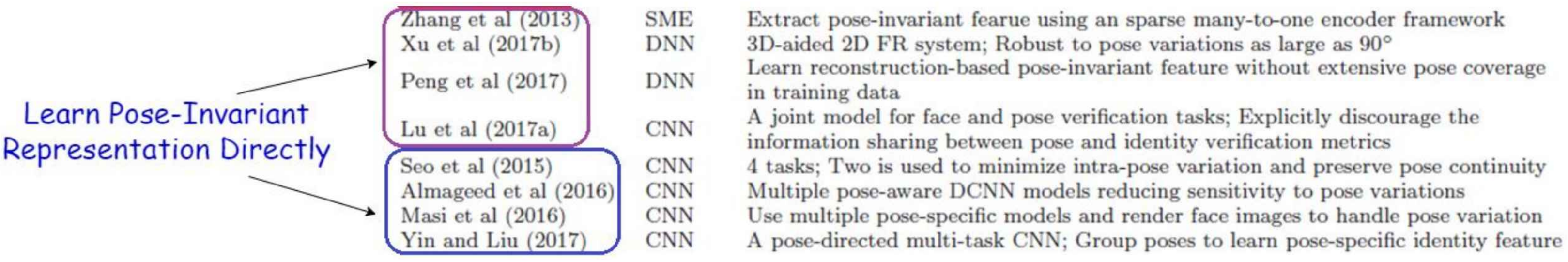


Hu L, Kan M, Shan S, Song X, Chen X (2017b) Ldf-net: Learning a displacement field network for face recognition across pose. In: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE, pp 9–16

Fig. 2. Schema of our method, LDF-Net. LDF-Net is an end-to-end method to learn the transformation from a non-frontal face image to a frontal one, composing of a displacement field network F and a translation layer T.

▪ Learn Pose-Invariant Representation Directly

- Focuses on learning a pose-invariant representation directly from non-frontal face images through:
  - either one joint model or
  - multiple pose-specific models



Learn Pose-Invariant Representation Directly

Zhang et al (2013)	SME	Extract pose-invariant fearue using an sparse many-to-one encoder framework
Xu et al (2017b)	DNN	3D-aided 2D FR system; Robust to pose variations as large as 90°
Peng et al (2017)	DNN	Learn reconstruction-based pose-invariant feature without extensive pose coverage in training data
Lu et al (2017a)	CNN	A joint model for face and pose verification tasks; Explicitly discourage the information sharing between pose and identity verification metrics
Seo et al (2015)	CNN	4 tasks; Two is used to minimize intra-pose variation and preserve pose continuity
Almageed et al (2016)	CNN	Multiple pose-aware DCNN models reducing sensitivity to pose variations
Masi et al (2016)	CNN	Use multiple pose-specific models and render face images to handle pose variation
Yin and Liu (2017)	CNN	A pose-directed multi-task CNN; Group poses to learn pose-specific identity feature

One joint model

- RF-SME (Zhang et al, 2013)
  - ✔ Random faces guided sparse many-to-one encoder
  - ✔ used a sparse many-to-one encoder to extract discriminative features

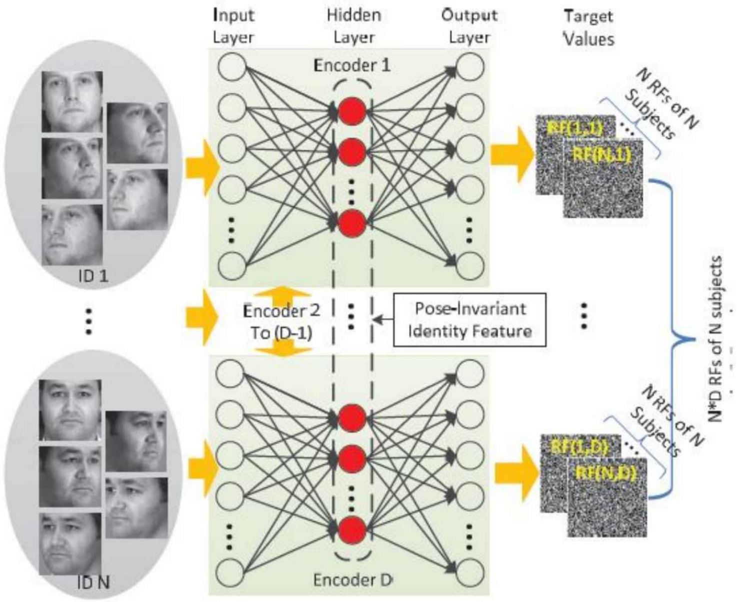
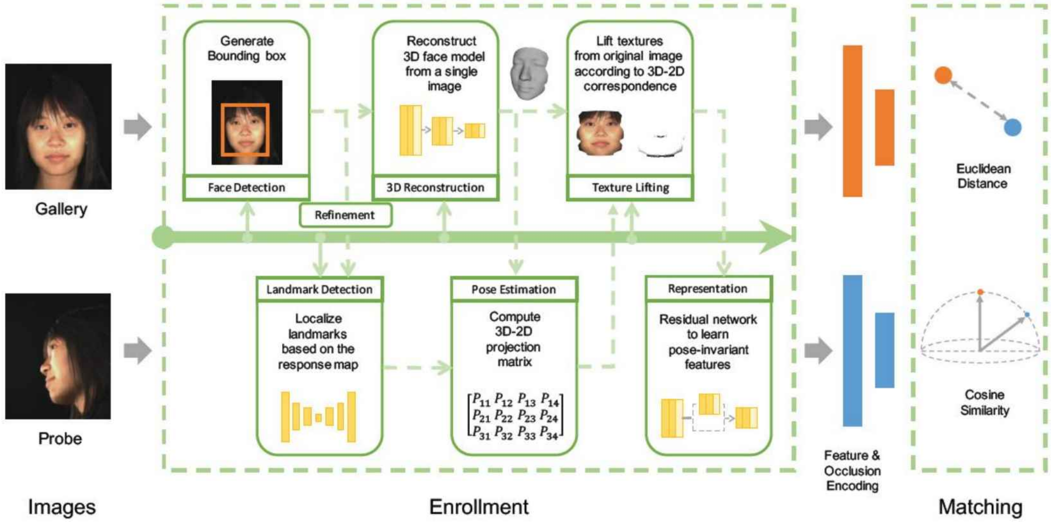


Figure 1. Framework of random faces guided sparse many-to-one encoder. Each unique "ID" has many facial images in different poses. We feed them into the single-hidden-layer neural network, i.e., the encoder, and set the target values to be random faces (RF). We design D encoders and therefore have D random faces for each ID. The concatenated nodes in hidden layers compose the high-level pose-invariant feature (red nodes in the dash area).

Zhang Y, Shao M, Wong EK, Fu Y (2013) Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In: Proceedings of the IEEE Intl. Conf. on Computer Vision, pp 2416–2423

□ Xu et al (2017b)

✔ proposed a 3D-aided 2D face recognition system



Gallery
Probe
Generate Bounding box
Face Detection
Reconstruct 3D face model from a single image
3D Reconstruction
Lift textures from original image according to 3D-2D correspondence
Texture Lifting
Refinement
Landmark Detection
Localize landmarks based on the response map
Pose Estimation
Compute 3D-2D projection matrix
$\begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \end{bmatrix}$
Representation
Residual network to learn pose-invariant features
Feature & Occlusion Encoding
Euclidean Distance
Cosine Similarity
Images
Enrollment
Matching

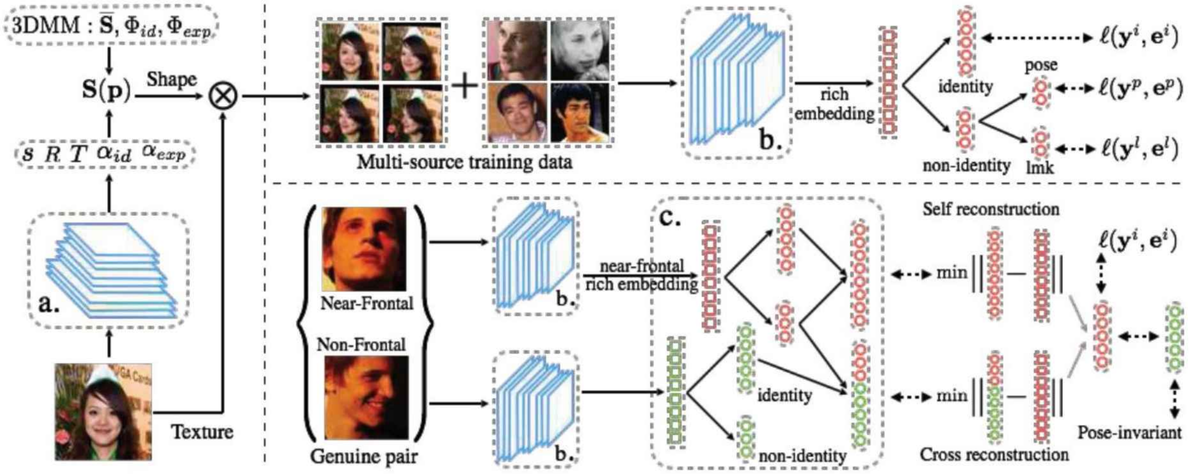
Figure 3: Depiction of the whole pipeline (follow the arrow in the middle) of 3D2D-PIFR. The rounded rectangles represent different modules. Dashed arrows represent the workflow. The enrollment encompasses the modules listed. A face first is detected then is transferred to localize landmarks. A 3D model is constructed directly from a 2D image with a bounding box. With 2D landmarks and a 3D model, a 3D-2D projection matrix can be estimated. The frontalized image and occlusion map are generated according to the 3D model and projection matrix. The pose robust features are extracted from these images along with occlusion encoding. The matching step computes features from visible parts and outputs a similarity score.

Xu X, Le H, Dou P, Wu Y, Kakadiaris IA (2017b) Evaluation of 3d-aided pose invariant 2d face recognition system. In: Proc. Intl. Joint Conf. on Biometrics, Denver, Colorado

□ Peng et al (2017)

✔ designed a reconstruction loss to regularize identity feature learning

✔ adopted a data driven synthesis strategy to enrich the diversity of poses



3DMM : $\bar{\mathbf{S}}, \Phi_{id}, \Phi_{exp}$
$\mathbf{S}(\mathbf{p})$
Shape
$s\ R\ T\ \alpha_{id}\ \alpha_{exp}$
a.
Texture
Multi-source training data
b.
rich embedding
identity
non-identity
pose
lmk
$\ell(\mathbf{y}^i, \mathbf{e}^i)$
$\ell(\mathbf{y}^p, \mathbf{e}^p)$
$\ell(\mathbf{y}^l, \mathbf{e}^l)$
Near-Frontal
Non-Frontal
Genuine pair
c.
near-frontal rich embedding
identity
non-identity
Self reconstruction
min
Cross reconstruction
$\ell(\mathbf{y}^i, \mathbf{e}^i)$
Pose-invariant

Figure 2. An overview of the proposed approach. (a) *Pose-variant face generation* utilizes a 3D facial model to synthesize new viewpoints from near-frontal faces. (b) *Rich feature embedding* is then achieved by jointly learning the identity and non-identity features using multi-source supervisions. (c) Finally, *Disentangling by reconstruction* is applied to distill the identity feature from the non-identity one for robust and pose-invariant representation.

Peng X, Yu X, Sohn K, Metaxas DN, Chandraker M (2017) Reconstruction-based disentanglement for pose-invariant face recognition. intervals 20:12

multiple pose-specific models

- Yin and Liu (2017)
  - ✔ used multi-task CNN to extract pose-robust face features

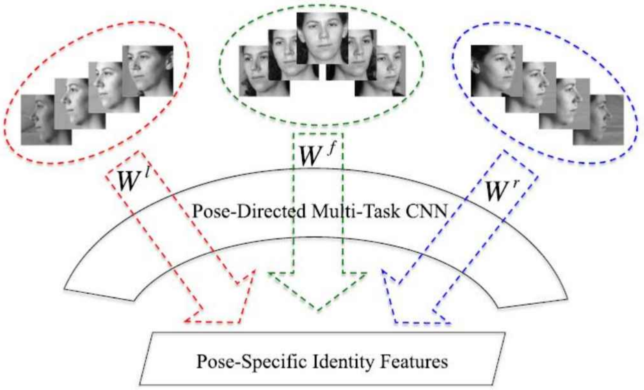


Fig. 4. The proposed pose-directed multi-task CNN aims to learn pose-specific identity features jointly for all pose groups.

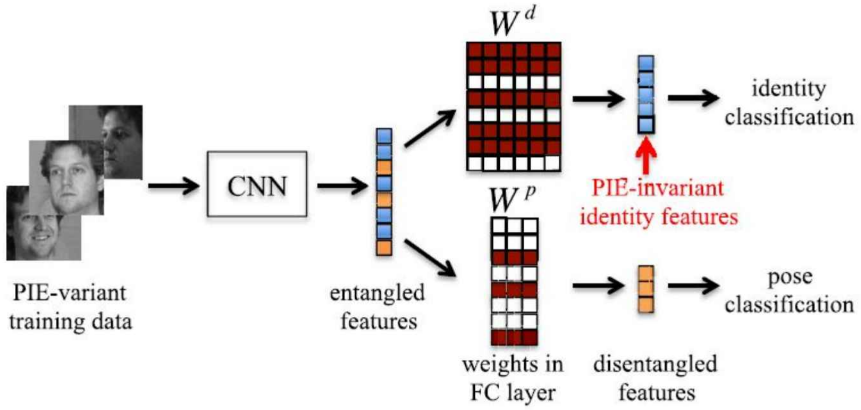


Fig. 1. We propose MTL for face recognition with identity classification as the main task and PIE classifications as the side tasks (only pose is illustrated in this figure for simplicity). A CNN framework learns entangled features from the data. The weight matrix in the fully connected layer of the main task is learnt to have close-to-zero values for PIE features in order to exclude PIE variations, which results in PIE-invariant identity features for face recognition.

Yin X, Liu X (2017) Multi-task convolutional neural network for pose-invariant face recognition. IEEE trans on Image Processing

▪ Synthesize a Frontal Face & Learn Pose-Invariant Representation Directly

- Tran et al (2017)
  - ✔ jointly merge face frontalization and pose-invariant identity representation learning
  - ✔ through a Disentangled Representation learning-GAN

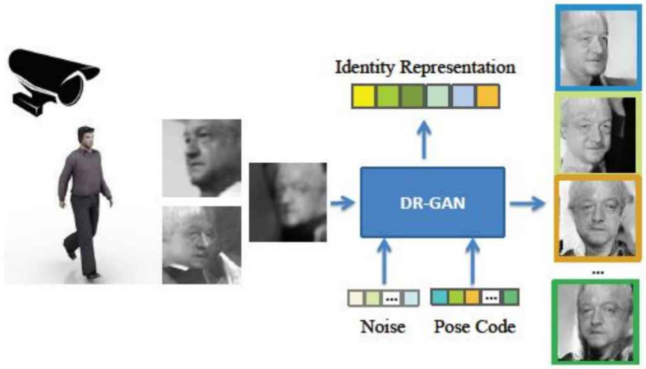
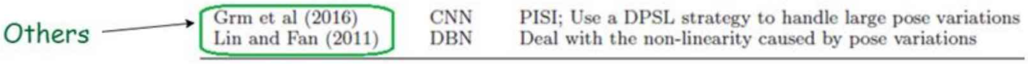


Figure 1: With one or multiple face images as the input, DR-GAN can produce an identity representation that is both discriminative and generative, i.e., the representation demonstrates superior PIFR performance, and can synthesize identity-preserving faces at target poses specified by the pose code.

Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: CVPR, vol 4, p 7

▪ Others



Grm et al (2016)	CNN	PISI; Use a DPSL strategy to handle large pose variations
Lin and Fan (2011)	DBN	Deal with the non-linearity caused by pose variations

□ Grm et al (2016)

✔ A value of index close to 1: same subject

✔ A value close to 0: different subjects

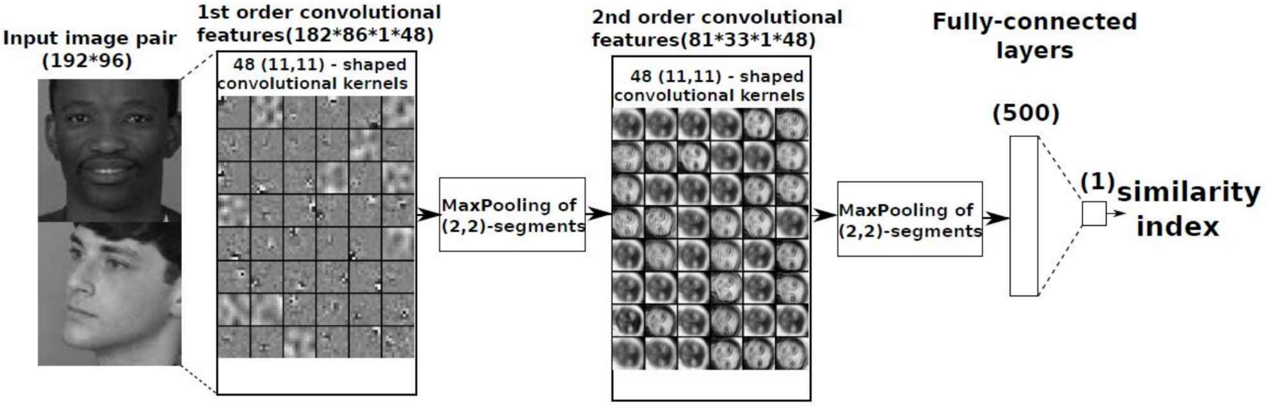


Fig. 3. The network architecture of the PISI (pose-invariant similarity index) model. The model exploits our DPSL strategy (deep pair-wise similarity learning) and takes two grayscale facial images with different poses as input and outputs a similarity index. The numbers in the brackets above the layers stand for the dimensionality of the layer outputs. Note that the outputs of the convolutional layers shown in the figure are only of an illustrative nature.

Grm K, Dobrisek S, Struc V (2016) Deep pair-wise similarity learning for face recognition. In: Biometrics and Forensics, Intl. Workshop on, IEEE, pp 1–6

❖ *Cross-Age Face recognition*

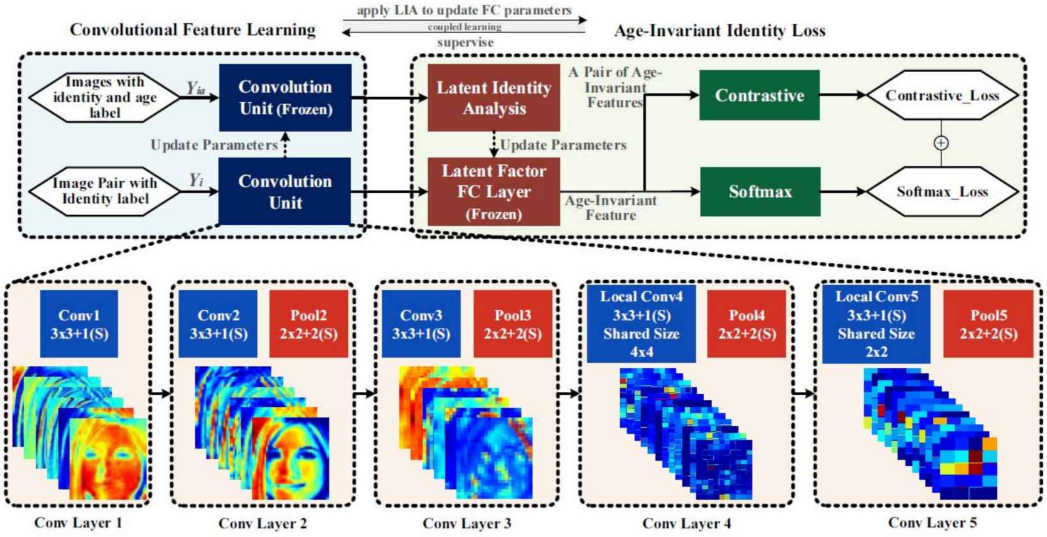
- Facial aging is also a challenge in FR
- With aging, the facial appearance can change significantly
- Some representative methods learn age-invariant features directly or indirectly

Table 11 Overview of deep learning methods for cross-age face recognition

Algorithm	Model	Description
Li et al (2015b)	CNN	Deep joint metric learning framework to learn age-invariant features
Wen et al (2016a)	CNN	A latent factor guided CNN; Construct latent identity analysis (LIA) module to help extract age-invariant features
Zheng et al (2017)	CNN	An age estimation task guided CNN; Learn age-invariant features on training data with age label and identity label
Xu et al (2017a)	AE	Coupled AE networks to handle age-invariant FR and retrieval problem
Antipov et al (2017b)	GAN	Age-cGAN; Synthesize aging/rejuvenation of the input face images to some predefined age categories to handle age variant
Antipov et al (2017a)	GAN	Resolve the issue that Age-cGAN cannot be directly used for improving face verification
Wang et al (2017d)	CNN	Cross-age FV by setting FV as primary learning task and age estimation as auxiliary learning task
Li et al (2018)	CNN	Present a distance metric optimization driven learning approach that integrates traditional steps via a DCNN

- ## Directly learn age-invariant features

- Wen et al (2016a)
  - ✔ A latent factor guided CNN;
  - ✔ Construct latent identity analysis (LIA) module to help extract age-invariant features



Wen Y, Li Z, Qiao Y (2016a) Latent factor guided convolutional neural networks for age-invariant face recognition. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 4893–4901

Figure 3. The architecture of the proposed LF-CNNs and its training process. *Frozen* layer only performs regular forward and backward calculations, but does not update their parameters (in other words, the parameters of this layers are fixed). The outside data $\boldsymbol{Y}_{\text{ia}}$ and the training data $\boldsymbol{Y}_{\text{ia}}$ are trained differently. Specifically, $\boldsymbol{Y}_{\text{ia}}$ and $\boldsymbol{Y}_{\text{ia}}$ are used for training the convolutional unit and the LF-FC layer respectively, following different pipelines. The *two parallel* convolution units are corresponding to a physical module in two stages (frozen and not frozen).

- Zheng et al (2017)
  - ✔ An age estimation task guided CNN
  - ✔ Learn age-invariant features on training data with age label and identity label

Zheng T, Deng W, Hu J (2017) Age estimation guided convolutional neural network for age-invariant face recognition. In: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, pp 1–9

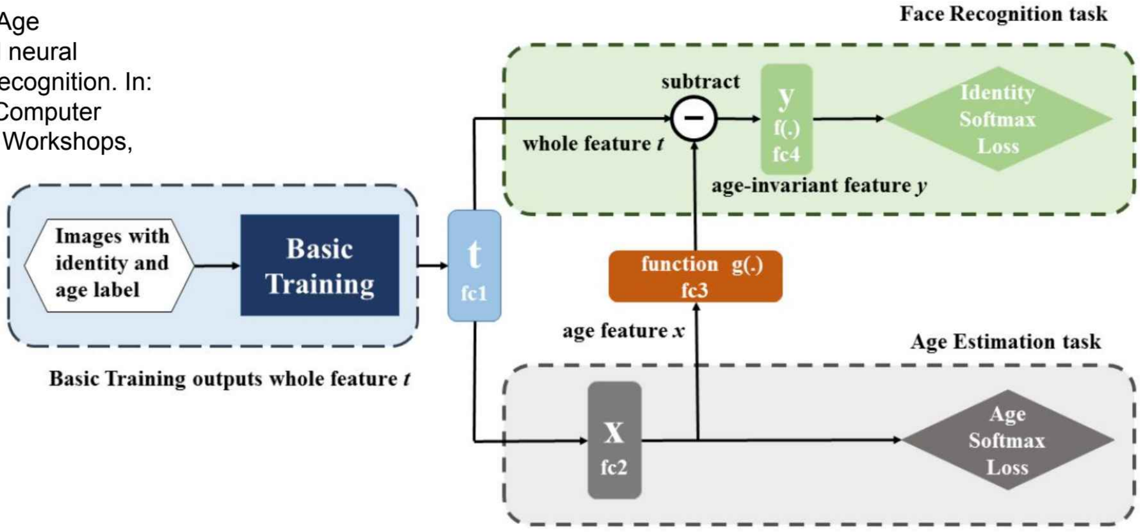


Figure 2. The architecture of the proposed AE-CNN. The formulation we use is $y = f(t - g(x))$ as shown in (1), t is the whole feature which contains age-related factor, x is the age feature obtained in age estimation task, y is the identity-specific feature for age-invariant face recognition, $g(.)$ is the function to obtain age factor which degrades the performance of face recognition from age feature. $f(.)$ is the function to better handle the relationship between the whole feature, age feature and identity-specific feature. The age estimation task and the face recognition task update parameters in the network at the same time.

 Wang et al (2017d)

- ✔ a multi-task deep neural network for cross-age face verification
- ✔ can effectively balance feature sharing and feature exclusion between face verification and age estimation, by exploiting an intrinsic, shared low-dimensional representation

Wang X, Zhou Y, Kong D, Currey J, Li D, Zhou J (2017d) Unleash the black magic in age: a multi-task deep neural network approach for cross-age face verification. In: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE, pp 596–603

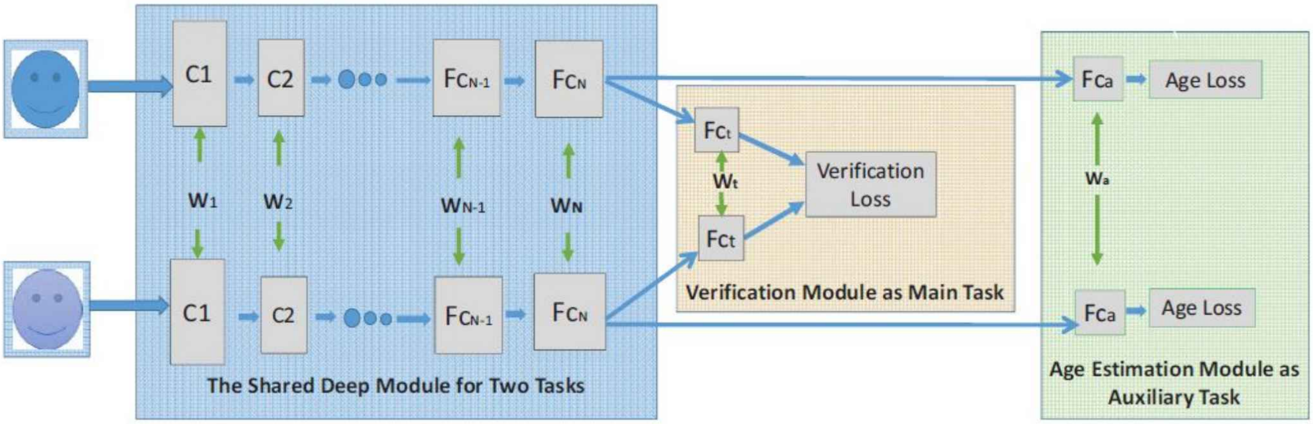

C1
C2
FcN-1
FcN
W1
W2
WN-1
WN
C1
C2
FcN-1
FcN
The Shared Deep Module for Two Tasks
Fct
Wt
Fct
Verification Loss
Verification Module as Main Task
Fca
Age Loss
Wa
Fca
Age Loss
Age Estimation Module as Auxiliary Task

Fig. 1. The proposed joint deep network architecture is based on the Siamese deep neural network for both face recognition and age estimation. It consists of two pipelines/tasks: (a) face recognition pipeline (b) age estimation pipeline. Given the pairwise images, the face recognition pipeline processes the features using deep neural network (DNN) encoded as $W^F = [W_1, W_2, \cdots, W_n]$ and feed them into the last layer encoded as W_t before minimizing the face recognition error given two pairwise images using a constractive loss (Eq.3). For each image, the age estimation pipeline processes the features using deep neural network encoded as $W^A = [W_1, W_2, \cdots, W_n]$ and feed them into the last layer encoded as W_a before minimizing the age estimation error for each image using a cross-entropy loss (Eq.4). In the figure, we assume the DNN structure is shared among the two pipelines/tasks, i.e., W^F23$= W^A$. In our framework, the layers $Fc_1 - Fc_N$ are directly adapted from VGG. Fc_N indicates Fc_7. Besides that, we designed the new layers Fc_t and Fc_a.

- Indirectly learn age-invariant features

 Antipov et al (2017b)

- ✔ proposed an Age-cGAN aging/rejuvenation method
- ✔ allowing to synthesize aging/rejuvenation of the input face images to some predefined age categories to handle age variant

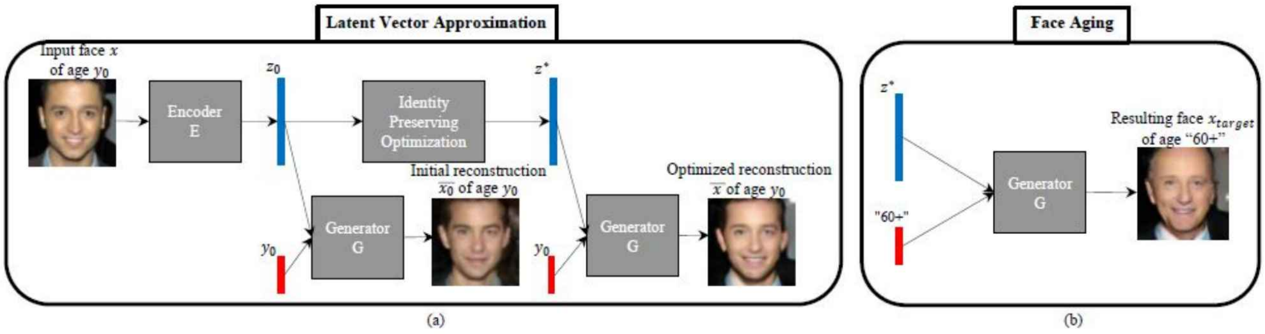
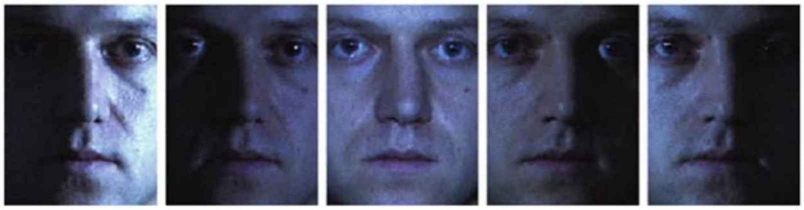

Latent Vector Approximation
Input face x of age y_0
Encoder E
z_0
Identity Preserving Optimization
z^*
Initial reconstruction $\overline{x_0}$ of age y_0
Generator G
y_0
Optimized reconstruction $\overline{x}$ of age y_0
Generator G
y_0
(a)
Face Aging
z^*
"60+"
Generator G
Resulting face x_{target} of age "60+"
(b)

Fig. 1. Our face aging method. (a) approximation of the latent vector to reconstruct the input image; (b) switching the age condition at the input of the generator G to perform face aging.

Antipov G, Baccouche M, Dugelay JL (2017b) Face aging with conditional generative adversarial networks. arXiv preprint arXiv:170201983

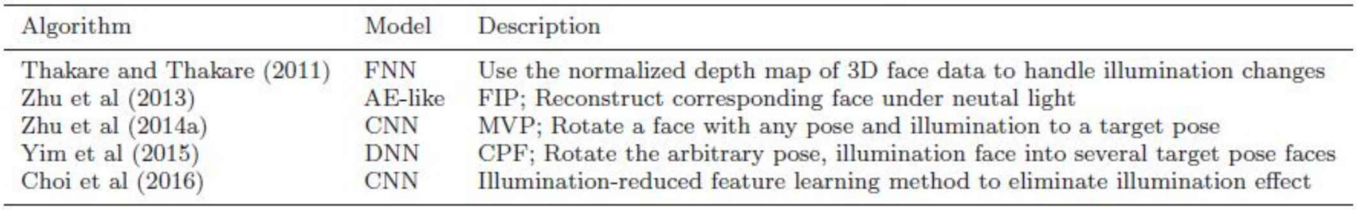
❖ *Illumination Changes*

- Lighting condition is also 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



- It is possible that the difference between two images of the same person taken under varying illumination to be greater than the difference between images of two different persons under the same illumination

Table 12 Overview of deep learning methods for dealing with illumination changes



Algorithm	Model	Description
Thakare and Thakare (2011)	FNN	Use the normalized depth map of 3D face data to handle illumination changes
Zhu et al (2013)	AE-like	FIP; Reconstruct corresponding face under neutal light
Zhu et al (2014a)	CNN	MVP; Rotate a face with any pose and illumination to a target pose
Yim et al (2015)	DNN	CPF; Rotate the arbitrary pose, illumination face into several target pose faces
Choi et al (2016)	CNN	Illumination-reduced feature learning method to eliminate illumination effect

□ Choi et al (2016)

✔ used a DCNN model to

- eliminate illumination effect
- maximize the discriminative power of feature representation

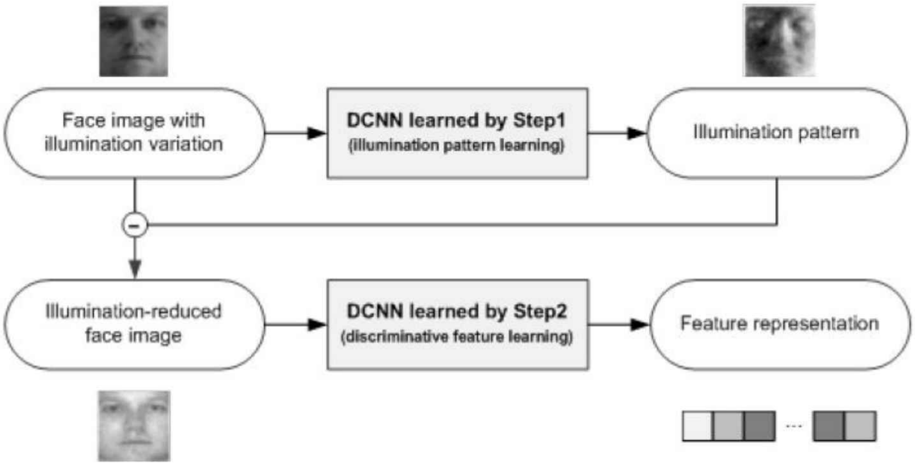

Face image with illumination variation
DCNN learned by Step1 (illumination pattern learning)
Illumination pattern
−
Illumination-reduced face image
DCNN learned by Step2 (discriminative feature learning)
Feature representation
···

Figure 1. FR method robust to illumination variations learned with the proposed two-step learning method.

Choi Y, Kim HI, Ro YM (2016) Two-step learning of deep convolutional neural network for discriminative face recognition under varying illumination. Electronic Imaging 2016(11):1–5

❖ *Facial Expression Variations*

- Facial expression changes may also impose problems for face recognition
- Facial deformations with expressions can change the appearance
- Researchers have used deep learning methods to address the expression problems

Pathirage et al (2015)

- ✔ a stacked denoising autoencoder for expression-robust feature acquisition
- ✔ exploits contributions of different color components in different local face regions
- ✔ by recovering the neutral expression from various expressions,
- ✔ and processes the faces with dynamic expressions progressively

Pathirage CSN, Li L, Liu W, Zhang M (2015) Stacked face de-noising auto encoders for expression-robust face recognition. In: Digital Image Computing: Techniques and Applications, Intl. Conf. on, IEEE, pp 1–8

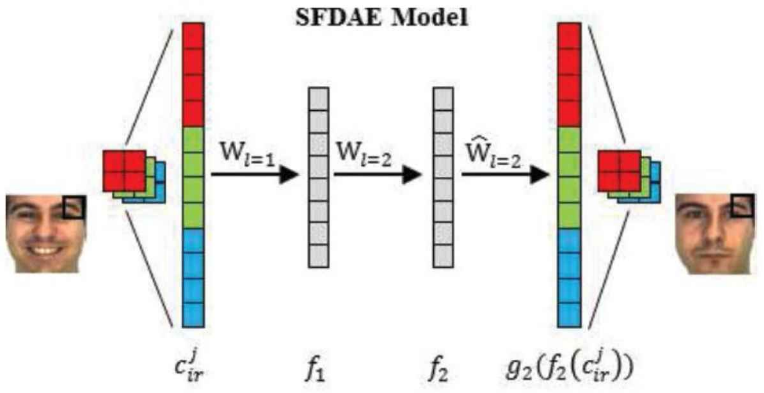


Fig. 4. The proposed SFDAE model where $f_1 \in \mathbb{R}^{50}$ denotes low dimensional noisy feature learnt at layer 1, while $f_2 \in \mathbb{R}^{50}$ denotes the noiseless feature learnt at layer 2 in the observed low dimensional space. We halves the image space by 50% to constraint the model to learn an effective low dimensional feature.

Liu et al (2016a)

- ✔ fused 2D images of a face and motion history images (MHIs), which are generated from the same face's image sequences with expressions to do face recognition

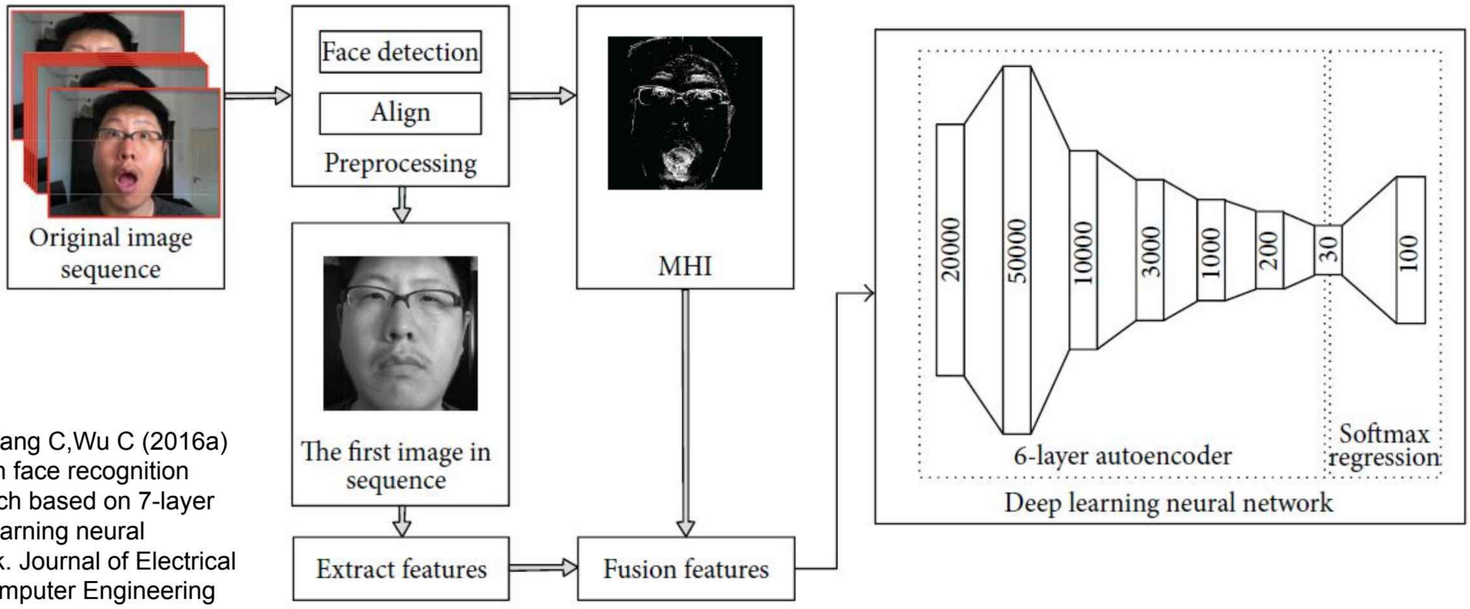


FIGURE 3: Structure of our network.

Liu J, Fang C,Wu C (2016a) A fusion face recognition approach based on 7-layer deep learning neural network. Journal of Electrical and Computer Engineering 2016

❖ *Facial Occlusions*

- The unavailability of the whole face is another challenge
- This happens when some parts of the face are missing or occluded, due to glasses, beard, moustache, scarf, etc.
- Such a problem can affect the recognition performance

Trigueros et al (2017)

✔ proposed a method to find out which parts of the face are more important to achieve a high recognition rate

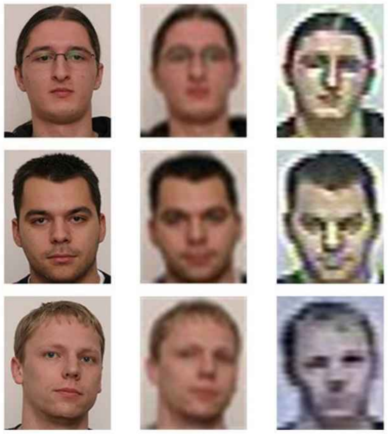
✔ and use that information during training to force the CNN to learn discriminative features from all face regions more equally,

✔ including those that typical approaches tend to pay less attention to

Trigueros DS, Meng L, Hartnett M (2017) Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. arXiv preprint arXiv:170707923

❖ *Low Resolution Face Images*

- Low resolution (LR) face images can degrade the face recognition performance significantly
  - Herrmann et al (2017)
    - ✔ compared the performance of three types of high-resolution CNN frameworks
    - ✔ on low-resolution face images
    - ✔ to search the most suitable one
      - o Microsoft’s residual architecture (He et al, 2016)
      - o Google’s inception architecture (Schroff et al, 2015)
      - o and classical VGGFace architecture (Parkhi et al, 2015)
    - ✔ found that the classical VGGFace architecture performs the best



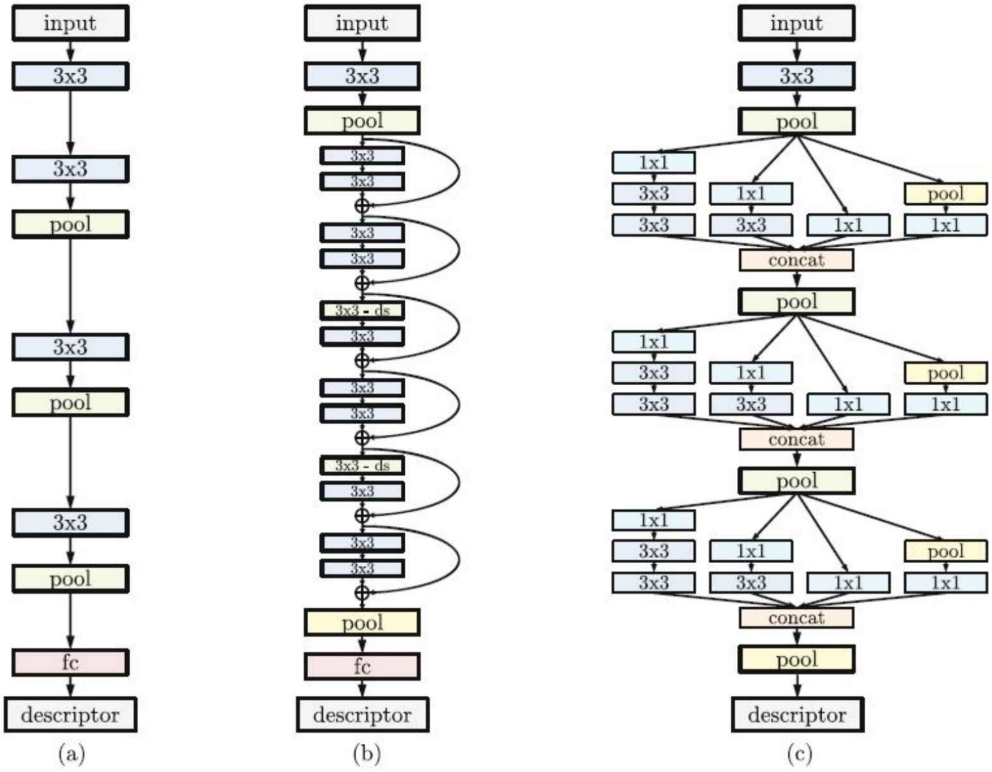


Fig. 1. Adapted LR networks for different architecture types: classical (a), residual (b) and inception (c). Green background denotes downsampling layers. (Color figure online)

❖ *Facial Makeup*

- Change the facial appearance
- Challenge the face recognition performance

Li et al (2017)

✔ proposed a bi-level adversarial network (BLAN) for makeup-invariant face verification

- Two adversarial networks are combined in an end-to-end deep network
  - with one in pixel level for reconstructing appealing facial images
  - the other in feature level for preserving the identity information

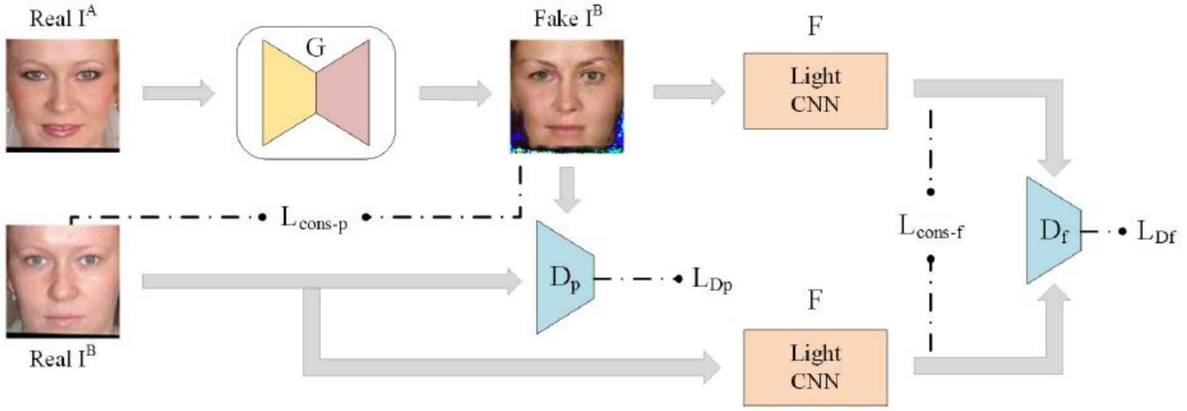


Figure 2: Diagram of the proposed Bi-level Adversarial Network. I^A is an input image with makeup while I^B stands for the corresponding non-makeup image. The generator G learns to fool the two discriminators, where D_p is on the pixel level and D_f on the feature level.

Li Y, Song L, Wu X, He R, Tan T (2017) Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. arXiv preprint arXiv:170903654

❖ *Mixed Variations*

- Deep learning methods are good at:
  - ✔ dealing with nonlinear characteristics in face images
  - ✔ and making the extracted features more discriminative
- In addition to focusing on one task
- there are a number of methods proposed to address more than one challenges

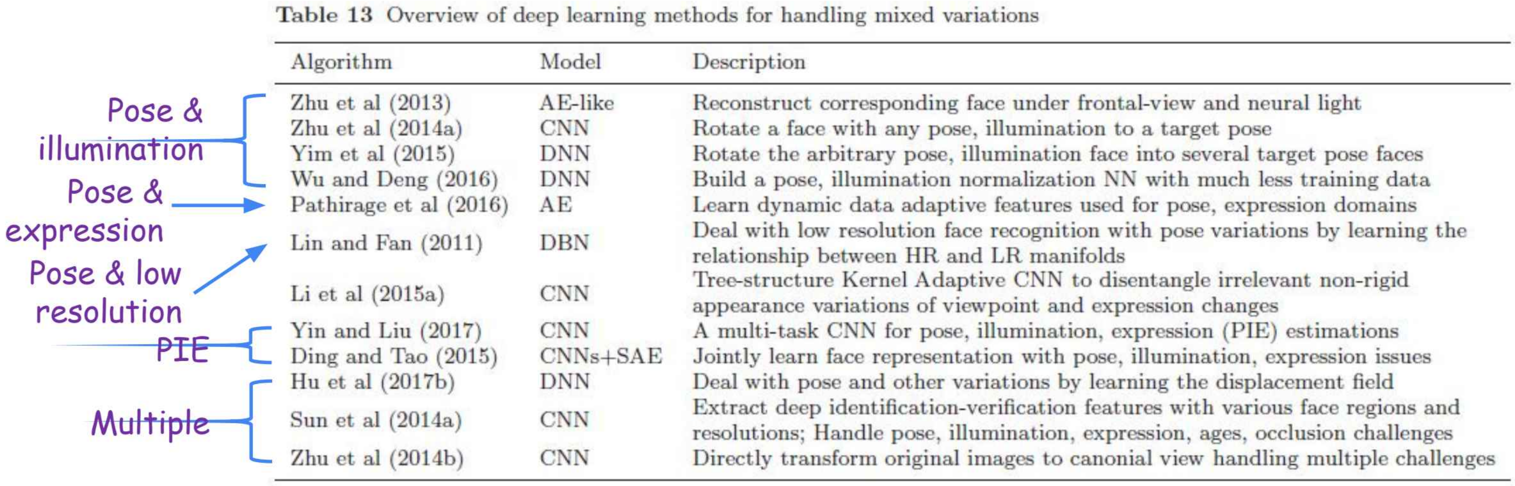


Table 13 Overview of deep learning methods for handling mixed variations

Algorithm	Model	Description
Zhu et al (2013)	AE-like	Reconstruct corresponding face under frontal-view and neural light
Zhu et al (2014a)	CNN	Rotate a face with any pose, illumination to a target pose
Yim et al (2015)	DNN	Rotate the arbitrary pose, illumination face into several target pose faces
Wu and Deng (2016)	DNN	Build a pose, illumination normalization NN with much less training data
Pathirage et al (2016)	AE	Learn dynamic data adaptive features used for pose, expression domains
Lin and Fan (2011)	DBN	Deal with low resolution face recognition with pose variations by learning the relationship between HR and LR manifolds
Li et al (2015a)	CNN	Tree-structure Kernel Adaptive CNN to disentangle irrelevant non-rigid appearance variations of viewpoint and expression changes
Yin and Liu (2017)	CNN	A multi-task CNN for pose, illumination, expression (PIE) estimations
Ding and Tao (2015)	CNNs+SAE	Jointly learn face representation with pose, illumination, expression issues
Hu et al (2017b)	DNN	Deal with pose and other variations by learning the displacement field
Sun et al (2014a)	CNN	Extract deep identification-verification features with various face regions and resolutions; Handle pose, illumination, expression, ages, occlusion challenges
Zhu et al (2014b)	CNN	Directly transform original images to canonial view handling multiple challenges

▪ Pose and illumination

□ Wu and Deng (2016)

- ✔ Build a pose and illumination normalization neural network with much less training data
- ✔ The idea is that:
  - o the output of normalization task should be identity-preserving
  - o and contains sufficient information of input identity to reconstruct the input image

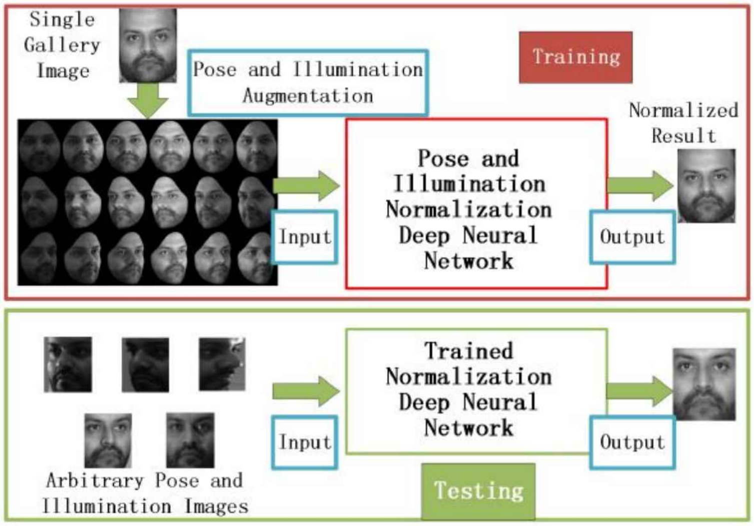

Single Gallery Image
Pose and Illumination Augmentation
Training
Normalized Result
Input
Pose and Illumination Normalization Deep Neural Network
Output
Trained Normalization Deep Neural Network
Input
Output
Arbitrary Pose and Illumination Images
Testing

Fig. 1. Visual illustration of proposed method.

Wu Z, Deng W (2016) One-shot deep neural network for pose and illumination normalization face recognition. In: Multimedia and Expo, Intl. Conf. on, IEEE, pp 1–6

- ✔ Use an auxiliary reconstruction task that reconstructs the original input image from the output of the normalization task,
- ✔ to improve the identity-preserving ability of the DNN

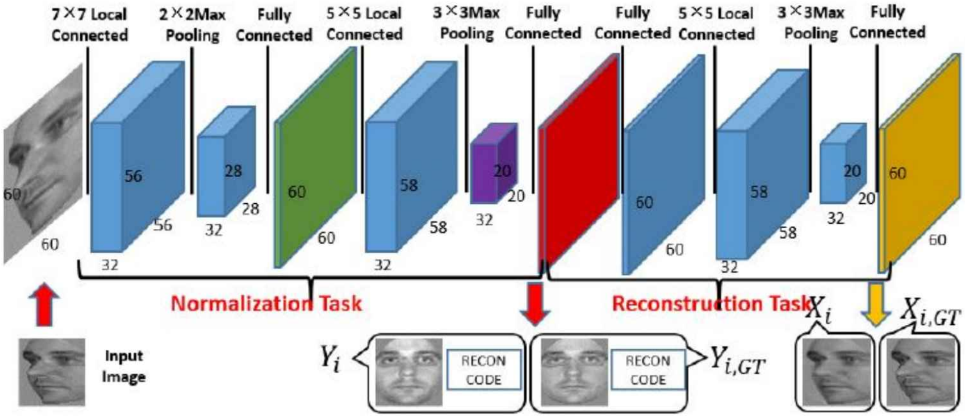


Fig. 5. Complete DNN architecture of our model.

▪ Pose and facial expression

□ Pathirage et al (2016):

✔ Deep Discriminant Analysis (DDA) Nets

✔ can learn dynamic data adaptive features used for various problems such as face pose and expressions

✔ consists of 3 interconnected learning processes:

- the progressive non-linear dimension reduction process:

  L1, L2; yield a low dimensional feature whose effective dimension is half the dimension of the original RGB features

- de-noising process

  L3; based on a strong supervisory signal which is the neutral frontal face

- Discrimination process

  L5; based on a single representative face image thus ensures the features observed in the reconstruction layer are highly discriminative

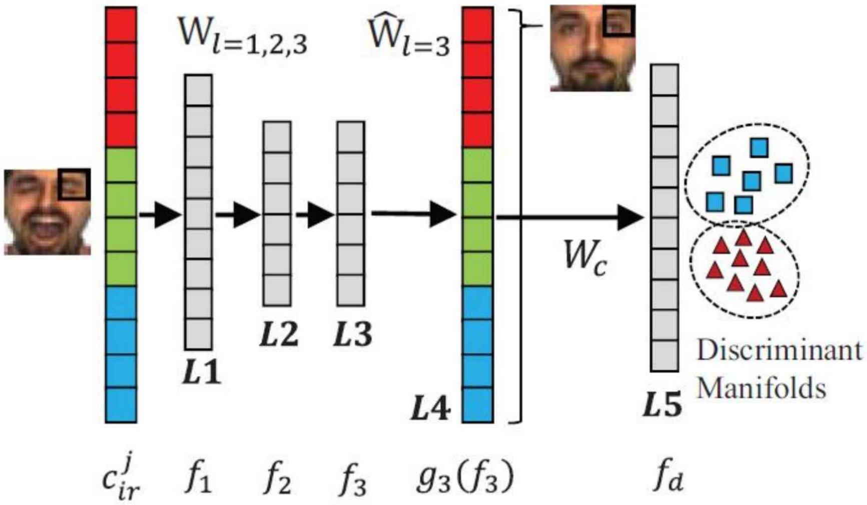


Fig. 1. DDA Net where $c_{ir}^{j} \in \mathbb{R}^{(36*3)}$, $f_1 \in \mathbb{R}^{75}$, $f_2 \in \mathbb{R}^{50}$ denote the combined patch feature and the low dimensional noisy feature learned at Layer 1 (L1) and Layer 2 (L2) respectively while $f_3 \in \mathbb{R}^{50}$ denotes the noise-less feature learned at Layer 3 (de-noising layer) in the observed low dimensional space. $g_3(.)$ represents the decoder function. Hence the discriminant layer where $f_d \in \mathbb{R}^{class\ count-1}$ is shown as the right most layer.

Pathirage CSN, Li L, Liu W (2016) Discriminant auto encoders for face recognition with expression and pose variations. In: Pattern Recognition, Intl. Conf. on, IEEE, pp 3512–3517

✔ In DDA, each shallow AE is trained to achieve simple but tractable goals required to address the global non-linear objective as a whole

✔ The framework follows a patch based approach to further refine the global non-linear objective into simpler tasks

✔ choose non-overlapping patches of the face image of size and stride 6x6 respectively

✔ it limits the number of parameters of the model that need to be learnt while training each DDA Net in a parallel environment

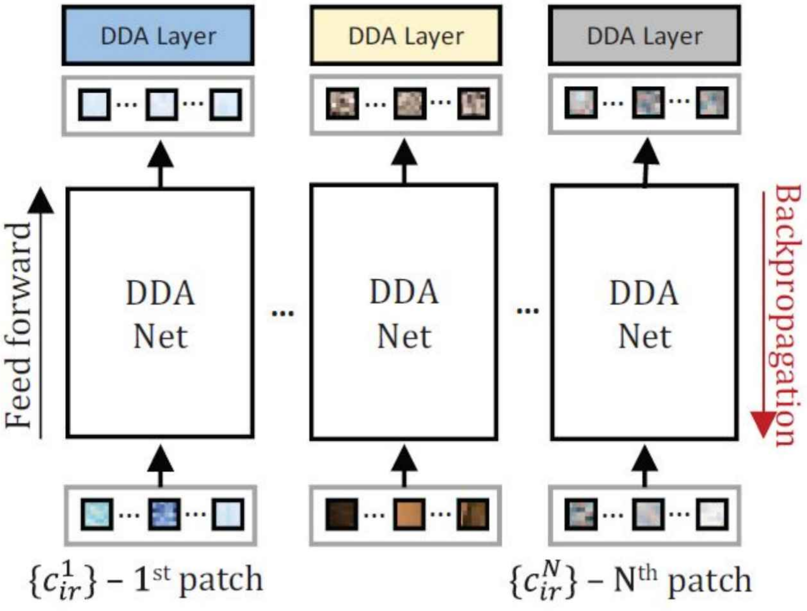


Fig. 2. Patch based DDA Framework that converts each patch of a face image to its corresponding frontal face patch followed by the non-linear discriminant analysis process (DDA layer).

▪ pose, illumination, expression (PIE)

□ Yin and Liu (2017)

- ✔ A multi-task CNN for pose, illumination, expression (PIE) estimations

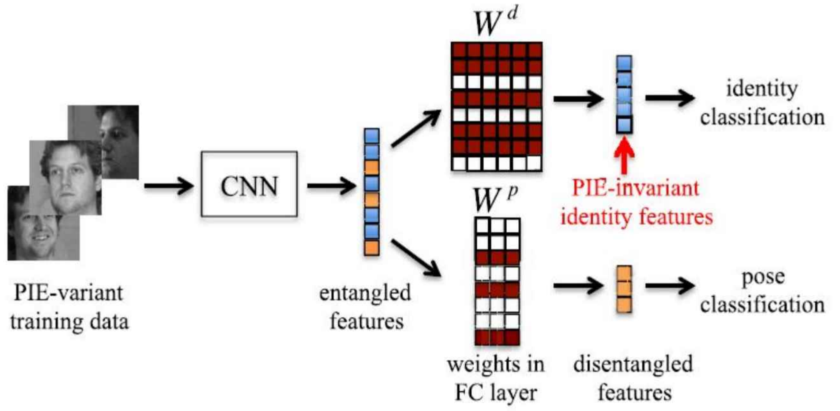


Fig. 1. We propose MTL for face recognition with identity classification as the main task and PIE classifications as the side tasks (only pose is illustrated in this figure for simplicity). A CNN framework learns entangled features from the data. The weight matrix in the fully connected layer of the main task is learnt to have close-to-zero values for PIE features in order to exclude PIE variations, which results in PIE-invariant identity features for face recognition.

Yin X, Liu X (2017) Multi-task convolutional neural network for pose-invariant face recognition. IEEE trans on Image Processing

▪ multiple challenges

□ Ding and Tao (2015)

- ✔ can jointly learn face representation with pose, illumination, expression issues by CNNs+SAE framework

□ DeepID2 (Sun et al, 2014a)

- ✔ can extract deep identification-verification features of images with various face regions and resolutions
- ✔ to deal with challenges including pose, illumination, expression, ages, occlusions

□ Zhu et al (2014b)

- ✔ proposed a deep learning framework
- ✔ can transform original images to a canonical view, which can also deal with other challenges

- Hu et al (2017b)

✔ proposed an end-to-end deep neural network
✔ used to transform a non-frontal face image into a frontal view by learning the displacement field,
✔ which reflects the shifting relationship of pixels from the non-frontal face image and the transformed frontal view

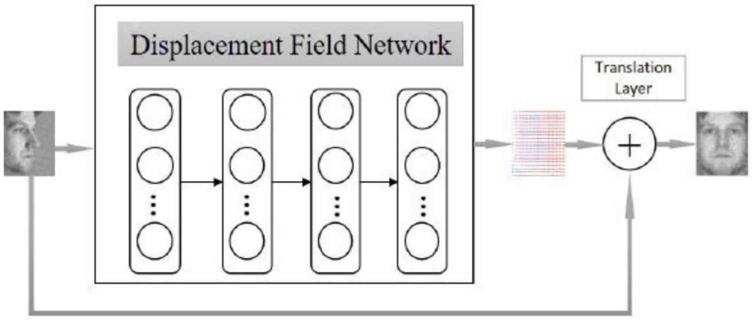


Fig. 2. Schema of our method, LDF-Net. LDF-Net is an end-to-end method to learn the transformation from a non-frontal face image to a frontal one, composing of a displacement field network F and a translation layer T.

Hu L, Kan M, Shan S, Song X, Chen X (2017b) Ldf-net: Learning a displacement field network for face recognition across pose. In: Automatic Face & Gesture Recognition, Intl. Conf. on, IEEE, pp 9–16

Video 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
- VFR aims to recognize whether a face video belongs to a certain subject
- A few algorithms have been developed to utilize varying approaches, ranging from frame by frame matching to advanced deep learning architectures
- The key issue is to build an appropriate visual representation of the video faces, such that it can effectively integrate the information across different frames together

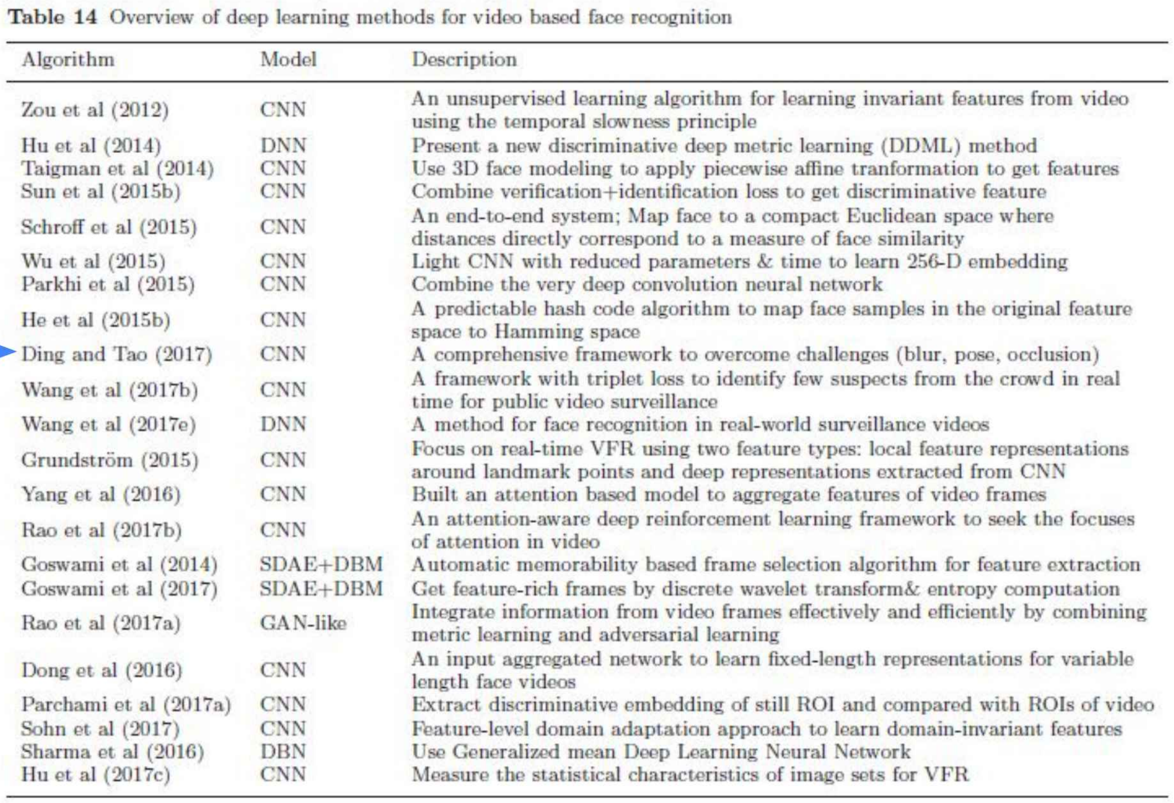


Table 14 Overview of deep learning methods for video based face recognition

Algorithm	Model	Description
Zou et al (2012)	CNN	An unsupervised learning algorithm for learning invariant features from video using the temporal slowness principle
Hu et al (2014)	DNN	Present a new discriminative deep metric learning (DDML) method
Taigman et al (2014)	CNN	Use 3D face modeling to apply piecewise affine tranformation to get features
Sun et al (2015b)	CNN	Combine verification+identification loss to get discriminative feature
Schroff et al (2015)	CNN	An end-to-end system; Map face to a compact Euclidean space where distances directly correspond to a measure of face similarity
Wu et al (2015)	CNN	Light CNN with reduced parameters & time to learn 256-D embedding
Parkhi et al (2015)	CNN	Combine the very deep convolution neural network
He et al (2015b)	CNN	A predictable hash code algorithm to map face samples in the original feature space to Hamming space
Ding and Tao (2017)	CNN	A comprehensive framework to overcome challenges (blur, pose, occlusion)
Wang et al (2017b)	CNN	A framework with triplet loss to identify few suspects from the crowd in real time for public video surveillance
Wang et al (2017e)	DNN	A method for face recognition in real-world surveillance videos
Grundström (2015)	CNN	Focus on real-time VFR using two feature types: local feature representations around landmark points and deep representations extracted from CNN
Yang et al (2016)	CNN	Built an attention based model to aggregate features of video frames
Rao et al (2017b)	CNN	An attention-aware deep reinforcement learning framework to seek the focuses of attention in video
Goswami et al (2014)	SDAE+DBM	Automatic memorability based frame selection algorithm for feature extraction
Goswami et al (2017)	SDAE+DBM	Get feature-rich frames by discrete wavelet transform& entropy computation
Rao et al (2017a)	GAN-like	Integrate information from video frames effectively and efficiently by combining metric learning and adversarial learning
Dong et al (2016)	CNN	An input aggregated network to learn fixed-length representations for variable length face videos
Parchami et al (2017a)	CNN	Extract discriminative embedding of still ROI and compared with ROIs of video
Sohn et al (2017)	CNN	Feature-level domain adaptation approach to learn domain-invariant features
Sharma et al (2016)	DBN	Use Generalized mean Deep Learning Neural Network
Hu et al (2017c)	CNN	Measure the statistical characteristics of image sets for VFR

Image &
Video

Address pose,
blur

Real-time video
surveillance

Vide
o

- Perform both image and video based FR

 DDML (Hu et al, 2014)

Hu J, Lu J, Tan YP (2014) Discriminative deep metric learning for face verification in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 1875–1882

 DeepFace (Taigman et al,2014)

Taigman Y, Yang M, Ranzato M, Wolf L (2014) Closing the gap to human-level performance in face verification. deepface. In: IEEE Computer Vision and Pattern Recognition

 DeepID2+ (Sun et al, 2015b)

Sun Y, Wang X, Tang X (2015b) Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 2892–2900

 FaceNet (Schroff et al, 2015)

Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 815–823

 Light CNN (Wu et al, 2015)

Wu X, He R, Sun Z, Tan T (2015) A light cnn for deep face representation with noisy labels. arXiv preprint arXiv:151102683

 VGGFace (Parkhi et al, 2015)

Parkhi OM, Vedaldi A, Zisserman A, et al (2015) Deep face recognition. In: BMVC, vol 1, p 6

 He et al (2015b), etc.

He R, Cai Y, Tan T, Davis L (2015b) Learning predictable binary codes for face indexing. Pattern Recognition 48(10):3160–3168

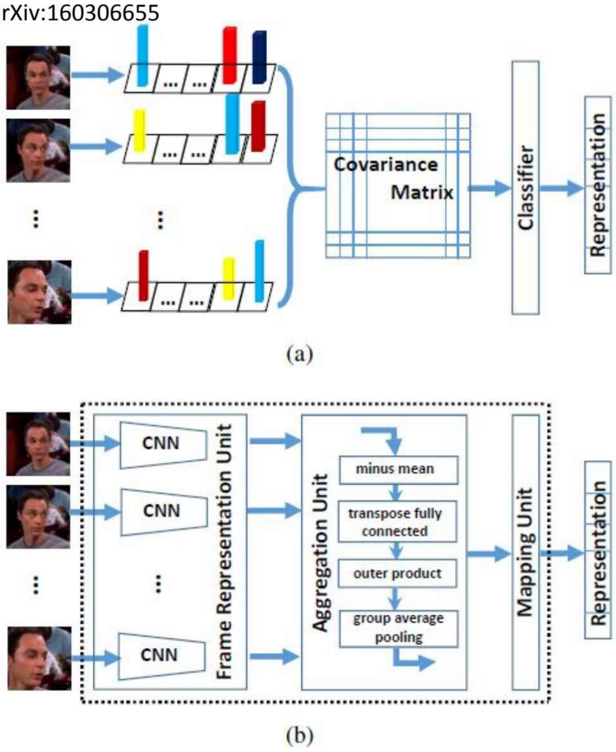
▪ Specially targeted on VFR

- The works mentioned above usually use face images as input
- To take full advantage of the useful information contained in videos, some methods learn face video representations directly

- Generally speaking, existing algorithms:
  - ✔ either select a small number of frames from all of the available frames
  - ✔ or use all frames to extract information-rich features

Dong et al (2016)

Dong Z, Jia S, Zhang C, Pei M (2016) Input aggregated network for face video representation. arXiv preprint arXiv:160306655



(a)

(b)

- ✔ To represent a face video, three steps are required:
  - ○ represent each face frame
  - ○ model the video clip
  - ○ map the video representation for the specific task
- ✔ Corresponding to the three steps, the input aggregated network contains three units:
  - ○ frame representation unit
  - ○ aggregation unit
  - ○ mapping unit

Figure 1: The illustrations of the traditional face video recognition method and our method. The traditional method (a) has three uncorrelated steps: extracting frame features, modeling face video clip, and learning classifier. Only the classifier learning procedure treats the final recognition task as optimal principle. Differently, the input aggregated network (b) integrates frame representation unit, aggregation unit, and mapping unit into an end-to-end system to learn the mapping from face videos to representations, and all the units serve for the final task.

- ✔ The aggregation unit aims at modeling the variable-length frame representations as fixed-length Riemannian manifold points
- ✔ The architecture contains four layers:
  - ○ Minus mean layer
  - ○ fully connected layer
  - ○ outer product layer
  - ○ group average pooling layer

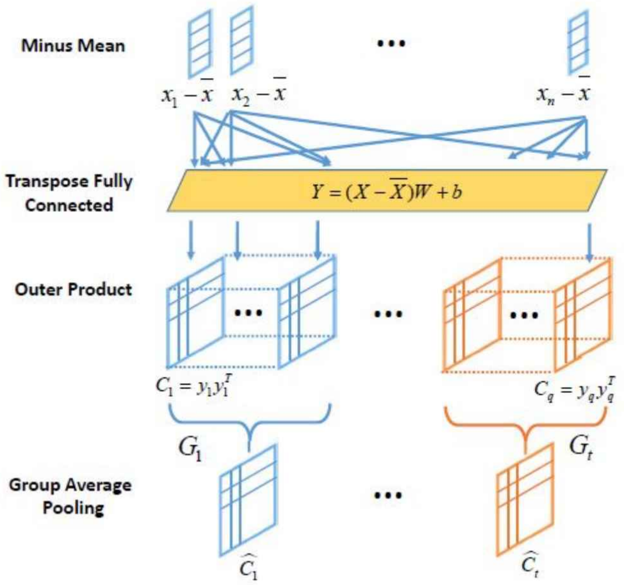


Figure 2: The architecture of the aggregation unit. The aggregation unit contains four layers: minus mean layer, transpose fully connected layer, outer product layer, and group average pooling layer.

- NAN (Yang et al, 2016)
  - ✔ Neural Aggregation Network
  - ✔ feature embedding module
    - ○ A CNN which maps each face frame into a feature representation
  - ✔ neural aggregation module
    - ○ composed of two content based attention blocks which is driven by a memory storing all the features extracted from the face video through the feature embedding module
    - ○ The output of the first attention block adapts the second, whose output is adopted as the aggregated representation of the video faces
    - ○ Due to the attention mechanism, this representation is invariant to the order of the face frames.

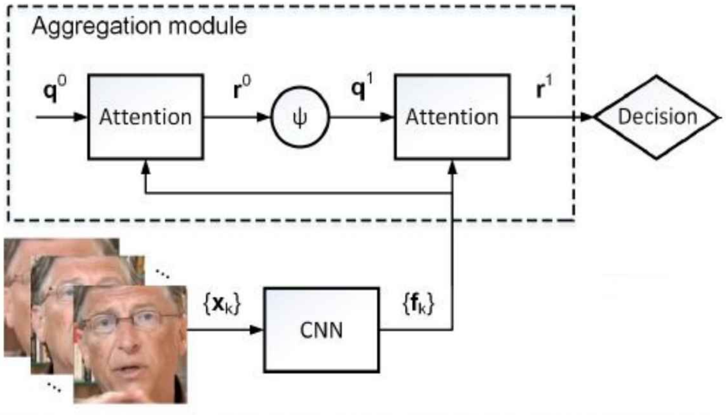


Figure 1. The face recognition framework of our method. All input faces $\{x_k\}$ are processed by a feature embedding module with a CNN, yielding a set of feature representations, $\{f_k\}$. These features are passed to the aggregation module, producing a 128-dimensional representation r^1 for the input video faces. This compact representation can then be used for the decision.

Yang J, Ren P, Chen D, Wen F, Li H, Hua G (2016) Neural aggregation network for video face recognition. arXiv preprint arXiv:160305474

○ Aggregation Module

- ✔ designed to take benefits from all frames in a video, potentially containing more discriminative information than a single image
- ✔ and handle arbitrary video size in an unified form, producing an order invariant representation

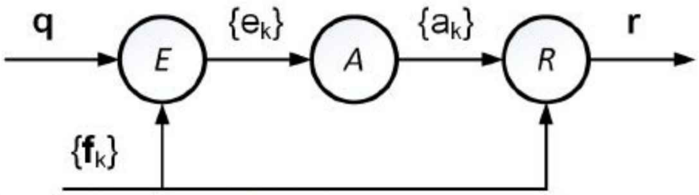


Figure 2. The attention block. It receives a set of feature vectors and filters each of them independently by a kernel q, yielding a set of scalars $\{e_k\}$. There scalars are then passed to a softmax operator, producing a set of weights $\{a_k\}$. Finally, the input feature vectors are fused via Eq. 1.

$$\mathbf{r} = \sum_k a_k \mathbf{f}_k. \quad (1)$$

□ Rao et al (2017a) ---See GAN

□ ADRL (Rao et al, 2017b)

- ✔ attention-aware deep reinforcement learning method
- ✔ aims to discard the misleading and confounding frames
- ✔ find the focuses of attentions in face videos for recognition

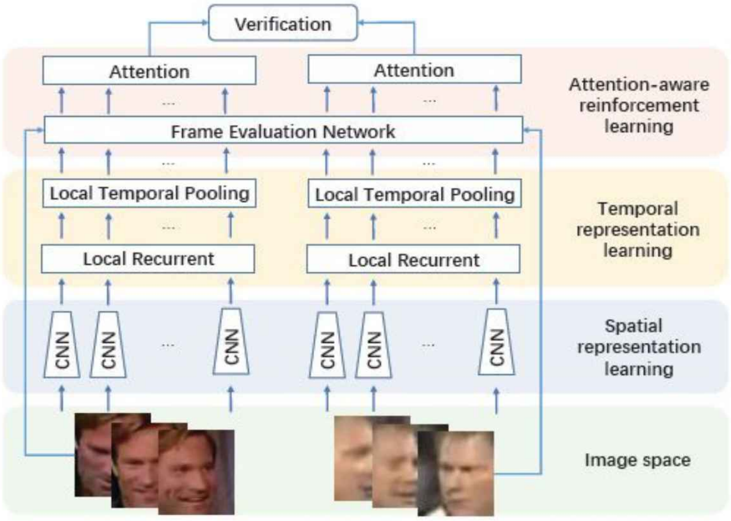


Figure 1. Flow-chart of our proposed method for video face recognition. Our approach takes a pair of face videos as the input and produces the temporal-spatial representations for each frame by using multiple stacked modules, including a convolutional neural network (CNN), a recurrent layer and a pooling layer with locality constraints, respectively. Then, a hard attention model with a frame evaluation network is trained by the proposed deep reinforcement learning method, which finds the attentions of the video pair for face verification.

Rao Y, Lu J, Zhou J (2017b) Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition, pp 3931–3940

- ✔ formulate the process of finding the attentions of videos as a Markov decision process
- ✔ Train the attention model through a deep reinforcement learning framework without using extra labels

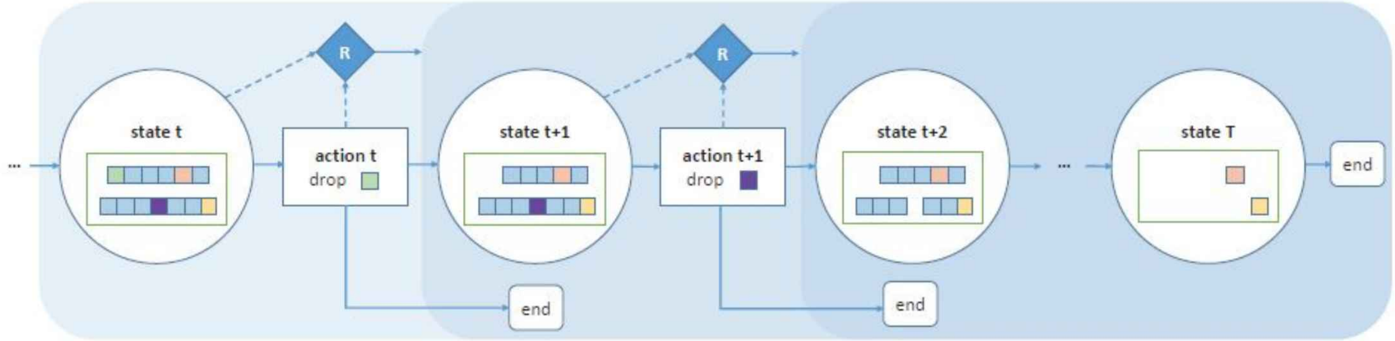


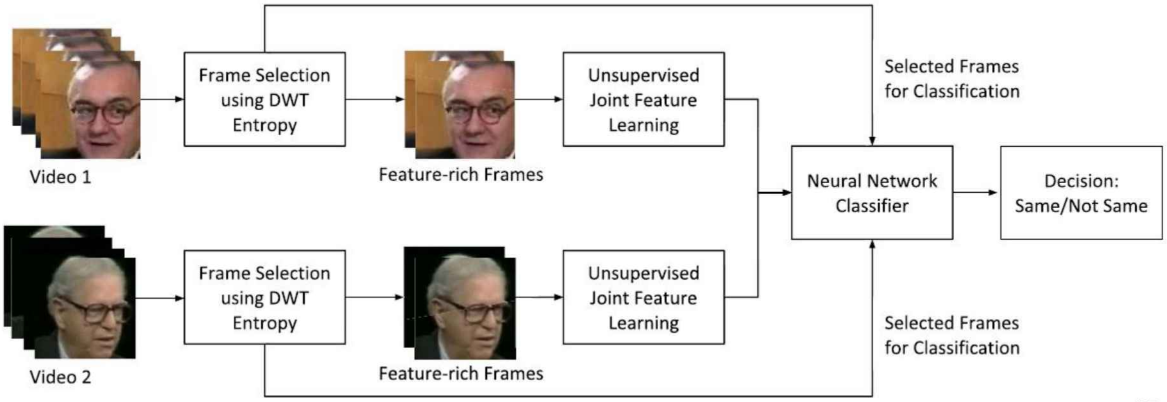
Figure 2. Markov decision process (MDP) of finding the focuses of attentions. States represent remaining frames after t steps, actions represent the decisions of dropping frames. Action a_t may lead to two states: state s_{t+1} and termination. Reward signal (R) is decided by the face recognition network C_1 depending on states and actions. States, actions, reward signals and terminations in MDP are illustrated by circles, rectangles, rhombuses and rounded rectangles, respectively.

- ✔ Unlike existing attention models, it takes information from both the image space and the feature space as the input to make better use of face information that is discarded in the feature learning process
- ✔ It is attention-aware, which seeks different attentions of videos for the recognition of different pairs of videos

Goswami et al (2017)

- ✔ select feature-rich frames from a video sequence using discrete wavelet transform and entropy computation
- ✔ followed by representation learning-based feature extraction:
  - o deep learning architecture
  - o a combination of stacked denoising sparse autoencoder (SDAE) and deep Boltzmann machine (DBM)
  - o formulation for joint representation in an autoencoder;
  - o update the loss function of DBM by including sparse and low rank regularization
- ✔ a multilayer neural network is used as the classifier to obtain the verification decision

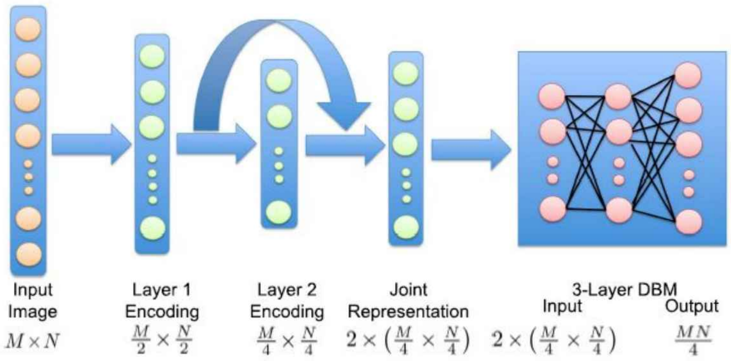
Goswami G, Vatsa M, Singh R (2017) Face verification via learned representation on feature-rich video frames. trans on Information Forensics and Security 12(7):1686–1698



Video 1
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification
Neural Network Classifier
Decision: Same/Not Same
Video 2
Frame Selection using DWT Entropy
Feature-rich Frames
Unsupervised Joint Feature Learning
Selected Frames for Classification

Fig. 3. Illustrating the steps involved in the proposed face recognition algorithm.

✔ representation learning-based feature extraction



Input Image
$M \times N$
Layer 1 Encoding
$\frac{M}{2} \times \frac{N}{2}$
Layer 2 Encoding
$\frac{M}{4} \times \frac{N}{4}$
Joint Representation
$2 \times (\frac{M}{4} \times \frac{N}{4})$
3-Layer DBM
Input
$2 \times (\frac{M}{4} \times \frac{N}{4})$
Output
$\frac{MN}{4}$

Fig. 5. Proposed deep learning architecture for facial representation: from input layer (image), two hidden layer representations are computed using SDAE encoding function. A joint representation is then obtained which combines the information from two SDAE encoding layers. Using joint representation as input, a DBM is used for computing a final feature vector.

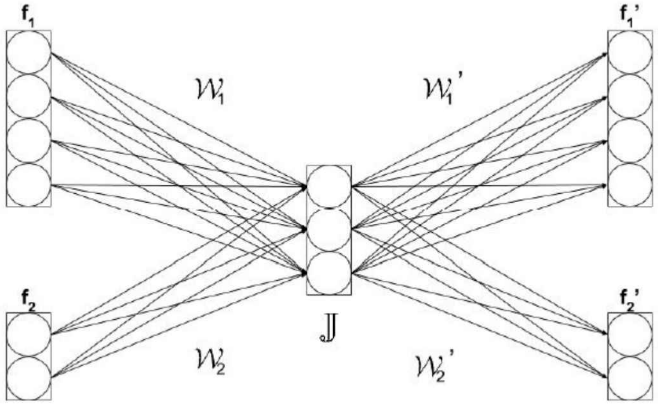


Fig. 6. Joint learning framework: features learned from the first and second levels of autoencoder, i.e., $\mathbf{f}_1$ and $\mathbf{f}_2$ are given as input to DBM to learn the joint representation $\mathbb{J}$.

- Face recognition in videos presents unique challenges due to the variations which can degrade the frame quality

□ TBE-CNN (Ding and Tao, 2017)

- ✔ Trunk-Branch Ensemble CNN
- ✔ A comprehensive framework based on CNN
- ✔ To address some challenges e.g., pose, occlusion and blur in VFR
- ✔ Trunk network implementation is based on GoogLeNet
- ✔ Divide the GoogLeNet layers into three levels:
  - o the low-level layers
  - o middle-level layers
  - o high-level layers

- ✔ The three layer levels successively extract features from the low- to the high-level
- ✔ Since low- and middle-level features represent local information, the trunk network and branch networks can share low- and middle-level layers
- ✔ In comparison, high-level features represent abstract and global information
- ✔ therefore, different models should have independent high-level layers

Ding C, Tao D (2017) Trunk-branch ensemble convolutional neural networks for video-based face recognition. IEEE trans on Pattern Analysis and Machine Intelligence

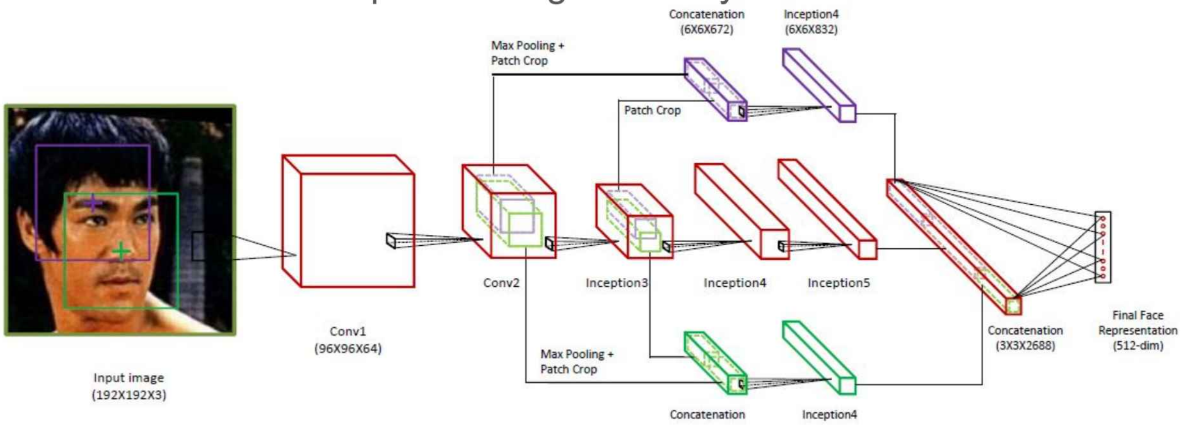


Fig. 3. Model architecture for Trunk-Branch Ensemble CNN (TBE-CNN). Note that a max pooling layer is omitted for simplicity following each convolution module, e.g., Conv1 and Inception 3. TBE-CNN is composed of one trunk network that learns representations for holistic face images and two branch networks that learn representations for image patches cropped around facial components. The trunk network and the branch networks share the same low- and middle-level layers, and they have individual high-level layers. The output feature maps of the trunk network and branch networks are fused by concatenation. The output of the last fully connected layer is utilized as the final face representation of one video frame.

Parchami M, Bashbaghi S, Granger E (2017a) Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In: Neural Networks, Intl. Joint Conf. on, IEEE, pp 4625–4632

- HaarNet (Parchami et al, 2017a)
  - ✔ Inspired by Ding and Tao (2017)
  - ✔ a novel end-to-end ensemble of DCNNs
  - ✔ to extract discriminative embedding of still regions of interest (ROI) and then compare it with regions of interests (ROIs) in video

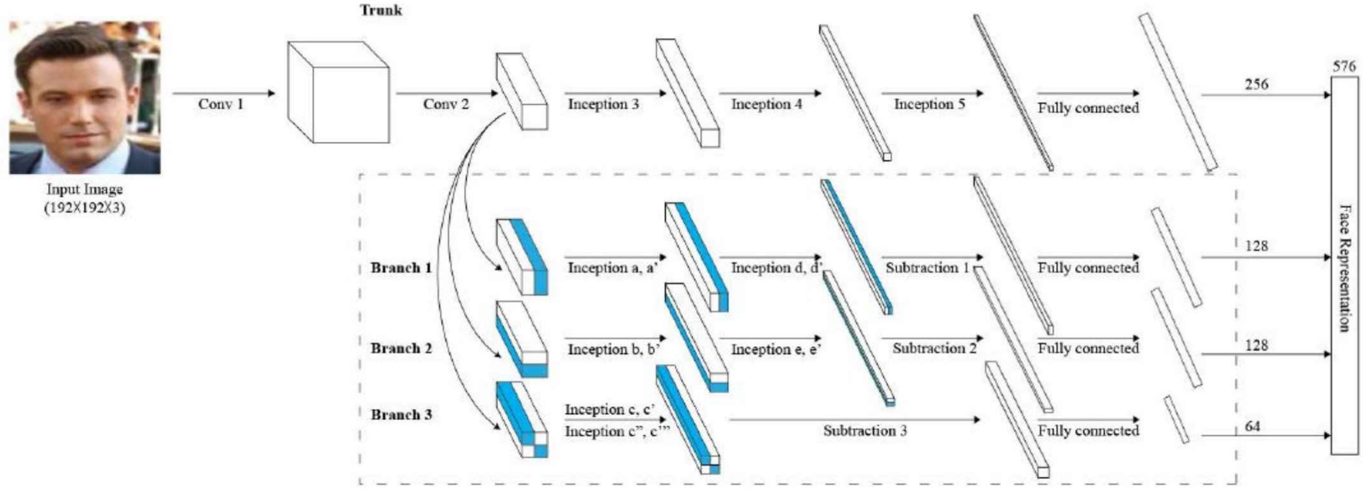


Fig. 2: HaarNet architecture for the trunk and three branches. (Max pooling layers after each inception and convolution layer are not shown for clarity).

- ✔ HaarNet uses a triplet-loss concept
- ✔ A batch of triplets composed of <anchor, positive, negative> is input to the architecture
- ✔ The output of the HaarNet is then *L*2 normalized prior to feed into the triplet-loss function in order to represent faces on a unit hyper-sphere

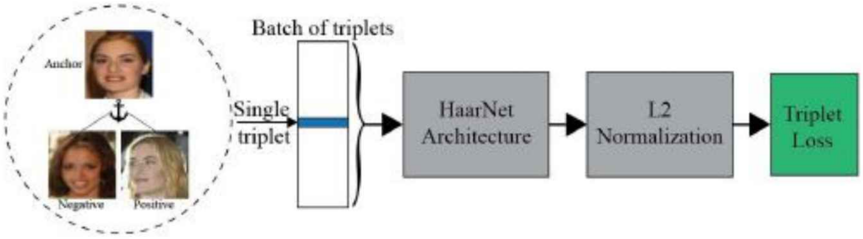


Fig. 3: Processing of triplets to compute the loss function. The network inputs a batch of triplets to the HaarNet architecture followed by an $L2$ Normalization.

- Sohn et al (2017)
  - ✔ an image to video feature-level domain adaptation approach to learn some domain-invariant discriminative representations for VFR
  - ✔ uses a pre-trained face recognition engine on labeled still images to extract discriminative information
  - ✔ adapts them to video domain by synthetic data augmentation
  - ✔ and then learns a domain-invariant feature through a domain adversarial discriminator

Sohn K, Liu S, Zhong G, Yu X, Yang MH, Chandraker M (2017) Unsupervised domain adaptation for face recognition in unlabeled videos. arXiv preprint arXiv:170802191

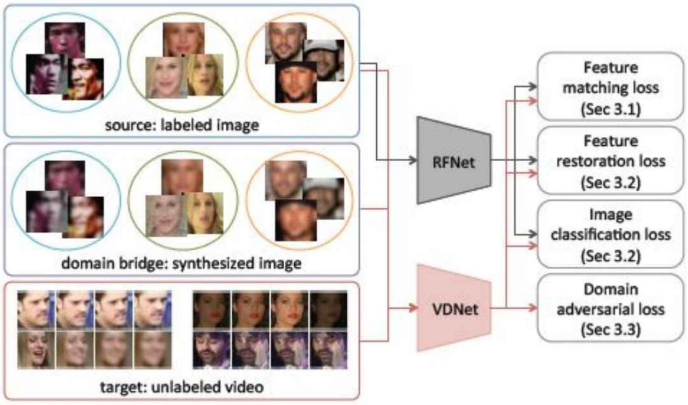

source: labeled image
domain bridge: synthesized image
target: unlabeled video
RFNet
VDNet
Feature matching loss (Sec 3.1)
Feature restoration loss (Sec 3.2)
Image classification loss (Sec 3.2)
Domain adversarial loss (Sec 3.3)

Figure 1. We propose an unsupervised domain adaptation method for video face recognition using large-scale unlabeled videos and labeled still images. To help bridge the gap between two domains, we introduce a new domain of synthesized images by applying a set of image transformations specific to videos such as motion blur to labeled images that simulates a video frame from still image. We utilize images, synthesized images, and unlabeled videos for domain adversarial training. Finally, we train a video domain-adapted network (VDNet) with domain adversarial loss (Section 3.3) as well as by distilling knowledge from pretrained reference network (RFNet) through feature matching (Section 3.1), feature restoration and image classification (Section 3.2) losses.

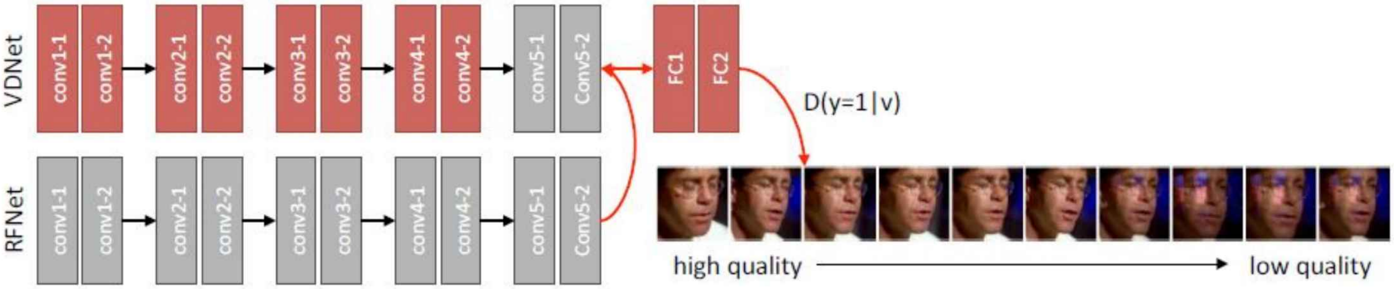


Figure 2. An illustration of network architecture for RFNet, VDNet and discriminator ($\mathcal{D}$). The red and gray blocks denote the trainable and fixed modules, respectively. VDNet not only shares the network architecture with RFNet, but also is initialized with the same network parameters. Once trained, $\mathcal{D}$ can sort the frames in a video sequence by indicating whether a frame is similar to images compatible to a face recognition engine and rejects those frames that are extremely ill-suited for face recognition.

□ ASML (Hu et al, 2017c)

✔ an Attention-Set based Metric Learning method

✔ to measure the statistical characteristics of image sets for VFR

- Handle real-world or real-time video surveillance

□ Wang et al (2017b)

✔ built a DCNN framework with a triplet supervisory signal

✔ to identify few suspects from the crowd in real time for public video surveillance

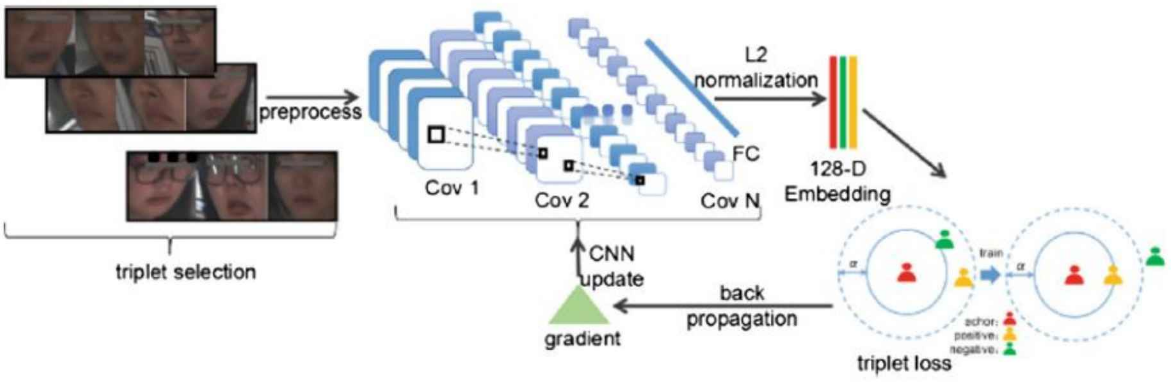


Fig. 2. End to end deep embedding training with triplet loss

Wang G, Sun Y, Geng K, Li S, Chen W (2017b) Deep embedding for face recognition in public video surveillance. In: Chinese Conf. on Biometric Recognition, Springer, pp 31–39

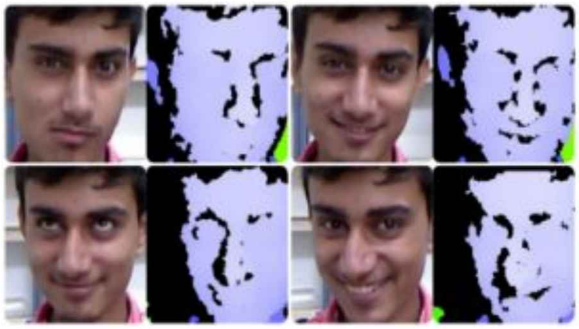
- Quality challenges, it is still not solved yet
- As videos usually contain many frames, it brings considerable computational burdens too

3D 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



- In recent years, some researches have focused on face recognition using 3D facial surface and shape

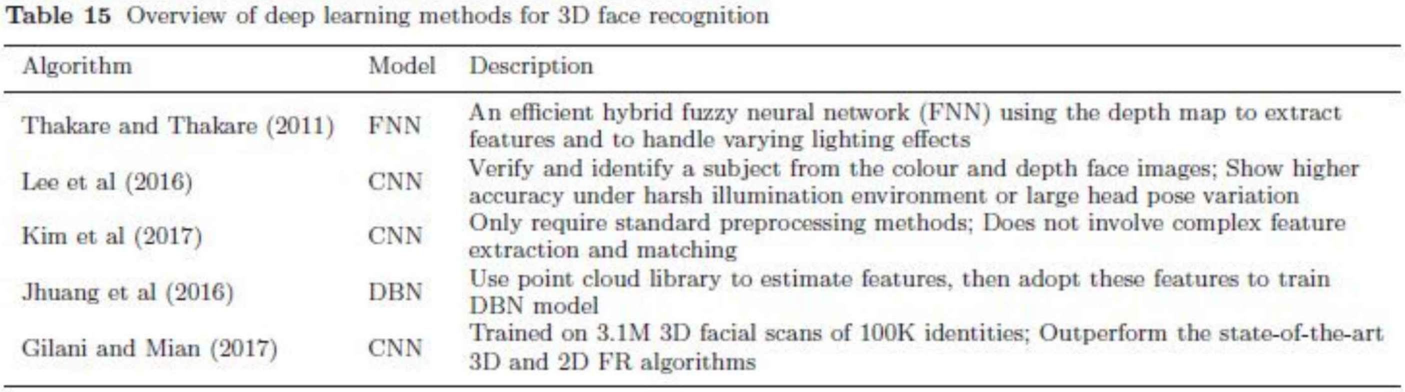


Table 15 Overview of deep learning methods for 3D face recognition

Algorithm	Model	Description
Thakare and Thakare (2011)	FNN	An efficient hybrid fuzzy neural network (FNN) using the depth map to extract features and to handle varying lighting effects
Lee et al (2016)	CNN	Verify and identify a subject from the colour and depth face images; Show higher accuracy under harsh illumination environment or large head pose variation
Kim et al (2017)	CNN	Only require standard preprocessing methods; Does not involve complex feature extraction and matching
Jhuang et al (2016)	DBN	Use point cloud library to estimate features, then adopt these features to train DBN model
Gilani and Mian (2017)	CNN	Trained on 3.1M 3D facial scans of 100K identities; Outperform the state-of-the-art 3D and 2D FR algorithms

Lee et al (2016)

- ✔ Face recognition from RGB-D images utilizes 2 complementary types of image data to achieve more accurate recognition
- ✔ 3 parts: (1) depth image recovery; (2) deep learning for feature extraction;
  (3) joint classification

To alleviate the problem of the limited size of available RGB-D data for deep learning

- ✔ firstly trained with color face dataset
- ✔ later fine-tuned on depth face images for transfer learning

Lee YC, Chen J, Tseng CW, Lai SH (2016) Accurate and robust face recognition from rgb-d images with a deep learning approach. In: BMVC

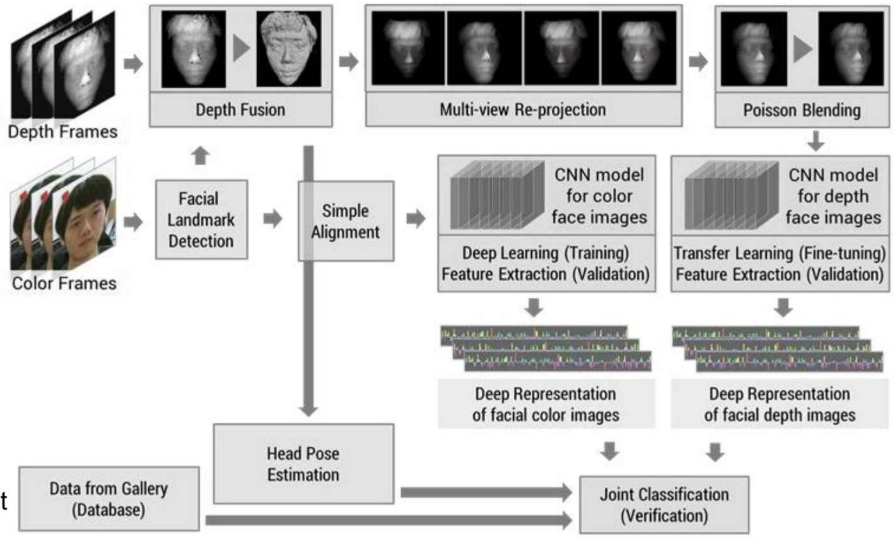

Depth Frames
Depth Fusion
Multi-view Re-projection
Poisson Blending
Color Frames
Facial Landmark Detection
Simple Alignment
CNN model for color face images
Deep Learning (Training) Feature Extraction (Validation)
CNN model for depth face images
Transfer Learning (Fine-tuning) Feature Extraction (Validation)
Deep Representation of facial color images
Deep Representation of facial depth images
Head Pose Estimation
Data from Gallery (Database)
Joint Classification (Verification)

Figure 1: A flowchart of our proposed system. Red region on depth maps means where depth information is lost.

FR3DNet (Gilani and Mian, 2017)

- ✔ A deep CNN model that is suited to 3D data
- ✔ Trained on 3.1Million 3D facial scans of 100K identities
- ✔ The skeleton architecture follows VGGFace but with a change in the conv layers

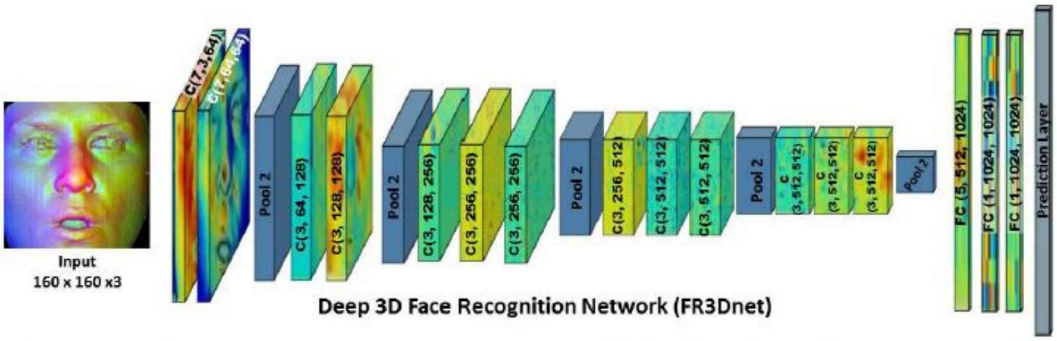


Figure 5. Architecture of our proposed *FR3DNet*. Every convolutional layer is followed by a rectifier layer.

Gilani SZ, Mian A (2017) Learning from millions of 3d scans for large-scale 3d face recognition. arXiv preprint arXiv:171105942

- Kim et al (2017)
  - ✔ 3D face recognition algorithm using a DCNN and a 3D augmentation technique

✔transfer learning from a CNN trained on 2D face images can effectively work for 3D face recognition by fine-tuning the CNN with a relatively small number of 3D facial scans

✔Propose a 3D face augmentation technique which synthesizes a number of different facial expressions from a single 3D face scan

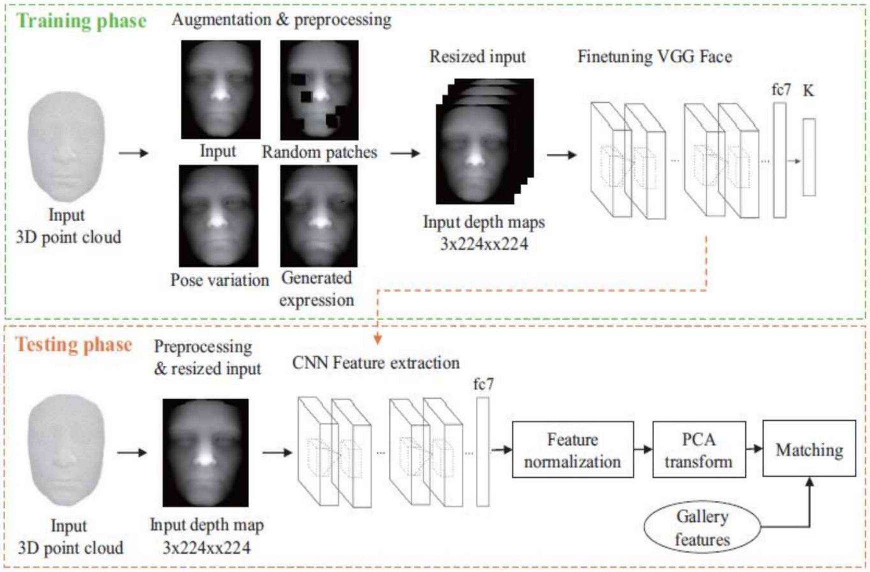


Figure 2: An overview of the proposed face identification system. In the training phase, we align 3D facial point clouds with a reference model, augment the point clouds, and convert them to 2D depth maps. Depth maps are resized to fit the size of VGG Face. In the testing phase, a probe scan is preprocessed and resized. Then, a face representation is extracted from the fine-tuned CNN. After normalization of features and Principal Component Analysis transform, one's identity is determined by the matching step.

Kim D, Hernandez M, Choi J, Medioni G (2017) Deep 3d face identification. arXiv preprint arXiv:170310714

- 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
  - as well as large test datasets
- Besides, the high cost of specialized 3D sensors limits their use for practical applications