A Survey on Deep Learning based Face Recognition

Na Zhang

Part IV: HFR

Heterogeneous Face Recognition (HFR)

- Problem of matching faces across different modalities
- Several specific HFR problems
  - VIS vs. Sketch
  - VIS vs. NIR
  - VIS vs. 3D
  - VIS vs. Video
  - Cross-Resolution
  - etc.
- The primary approaches
  - to extract common latent features between different modalities, so that a classifier trained on one modality may generalize to another

- An overview of generative HFR that can be used for face matching between different modalities

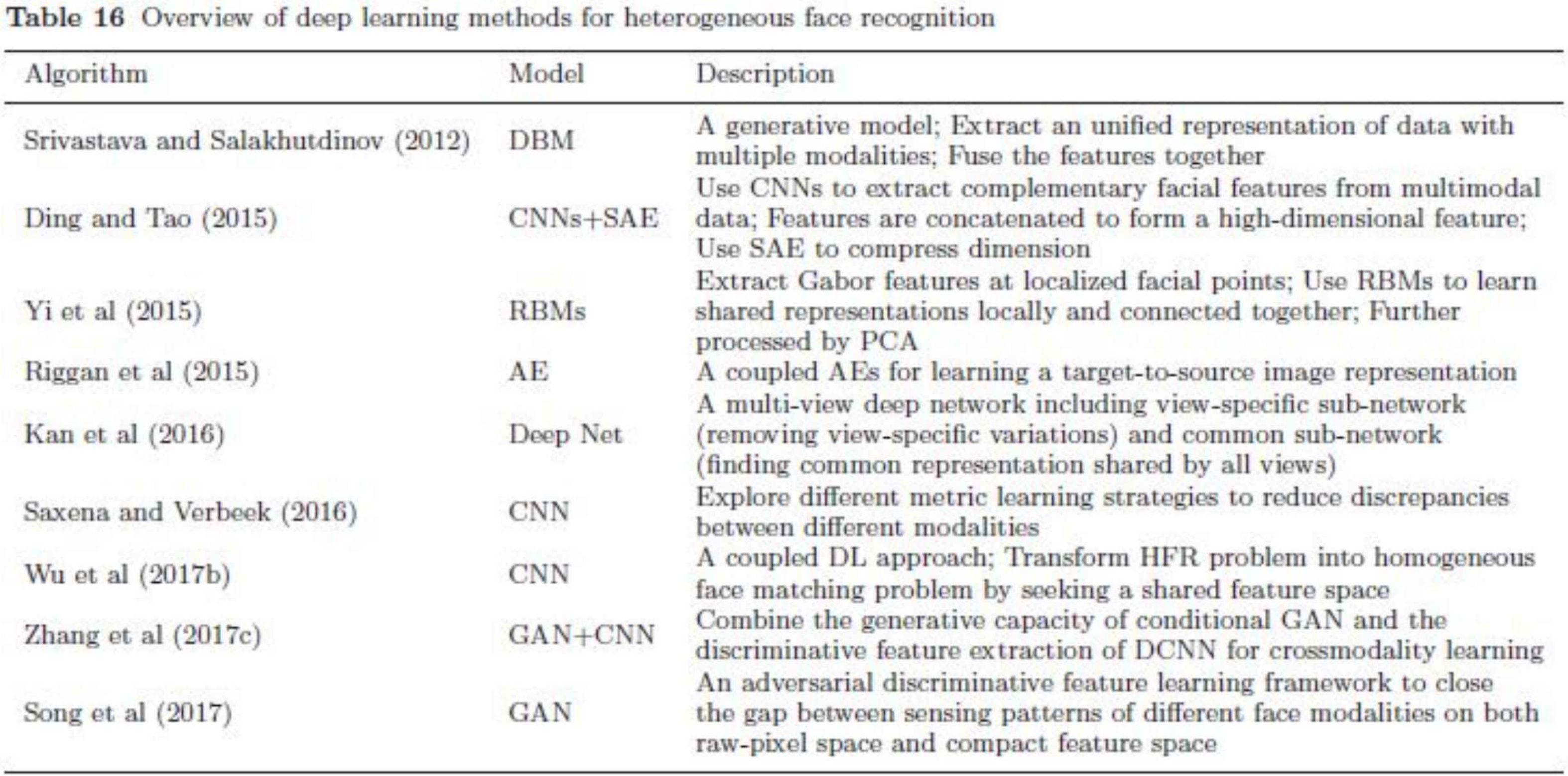


Table 16 Overview of deep learning methods for heterogeneous face recognition

Algorithm	Model	Description
Srivastava and Salakhutdinov (2012)	DBM	A generative model; Extract an unified representation of data with multiple modalities; Fuse the features together
Ding and Tao (2015)	CNNs+SAE	Use CNNs to extract complementary facial features from multimodal data; Features are concatenated to form a high-dimensional feature; Use SAE to compress dimension
Yi et al (2015)	RBMs	Extract Gabor features at localized facial points; Use RBMs to learn shared representations locally and connected together; Further processed by PCA
Riggan et al (2015)	AE	A coupled AEs for learning a target-to-source image representation
Kan et al (2016)	Deep Net	A multi-view deep network including view-specific sub-network (removing view-specific variations) and common sub-network (finding common representation shared by all views)
Saxena and Verbeek (2016)	CNN	Explore different metric learning strategies to reduce discrepancies between different modalities
Wu et al (2017b)	CNN	A coupled DL approach; Transform HFR problem into homogeneous face matching problem by seeking a shared feature space
Zhang et al (2017c)	GAN+CNN	Combine the generative capacity of conditional GAN and the discriminative feature extraction of DCNN for crossmodality learning
Song et al (2017)	GAN	An adversarial discriminative feature learning framework to close the gap between sensing patterns of different face modalities on both raw-pixel space and compact feature space

MM-DFR (Ding and Tao, 2015) --see Hybrid

- ✔ integrated a set of elaborately designed CNNs and a three-layer SAE
- ✔ The CNNs extract complementary facial features from multimodal data
- ✔ the extracted features are concatenated to form a high-dimensional feature vector, whose dimension is compressed by the SAE

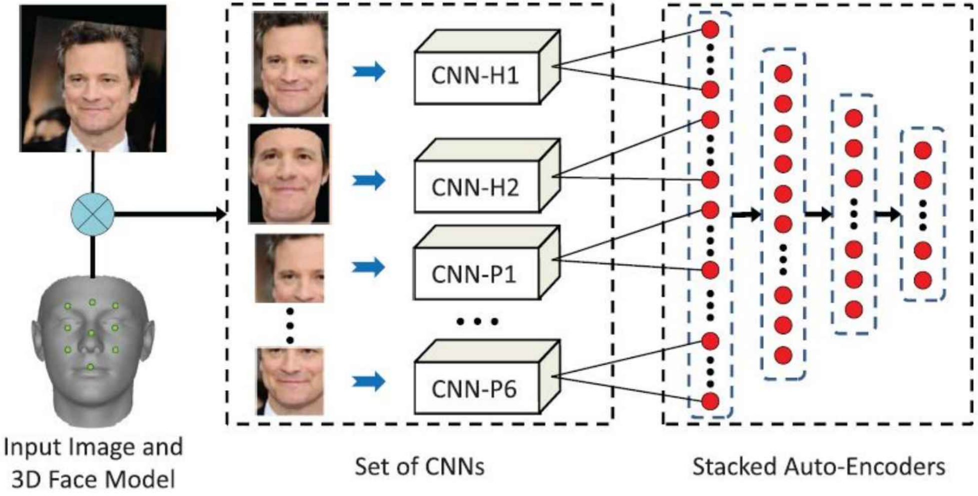


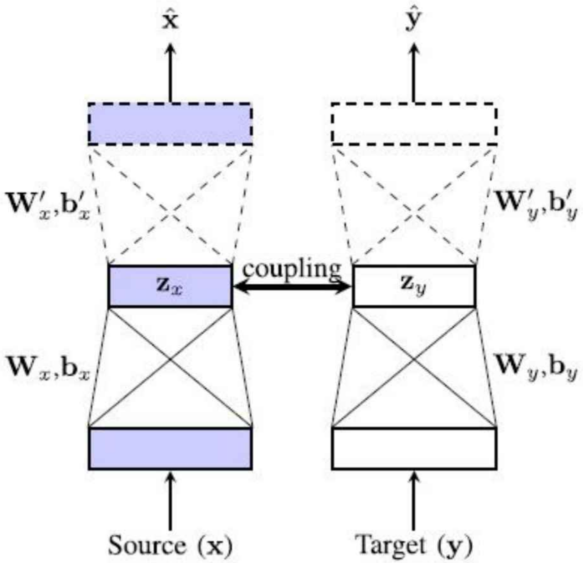
Fig. 2. Flowchart of the proposed multimodal deep face representation (MM-DFR) framework. MM-DFR is essentially composed of two steps: multimodal feature extraction using a set of CNNs and feature-level fusion of the set of CNN features using SAE.

Ding C, Tao D (2015) Robust face recognition via multimodal deep face representation. IEEE trans on Multimedia 17(11):2049–2058

Riggan et al (2015): ---see AE

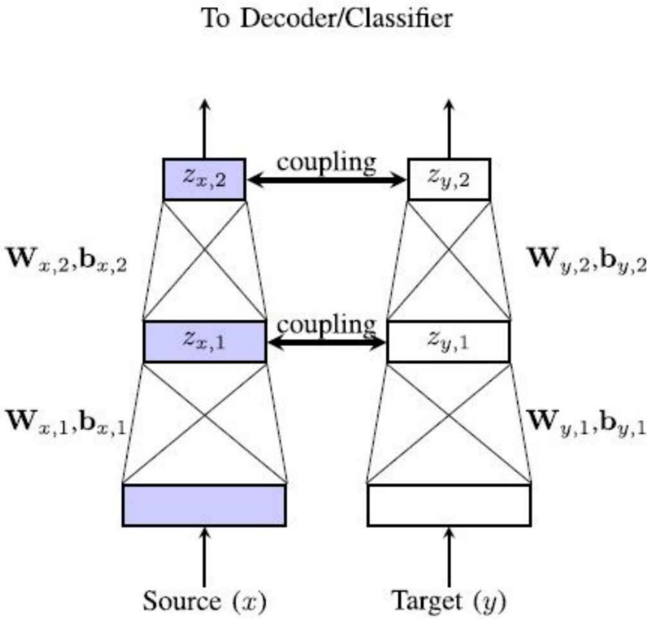
Riggan BS, Reale C, Nasrabadi NM (2015) Coupled auto-associative neural networks for heterogeneous face recognition. IEEE Access 3:1620–1632

- ✔ A coupled AEs for learning a target-to-source image representation
- ✔ A cross-modal transformation is learned by forcing the hidden units (latent features) of two neural networks to be as similar as possible, while simultaneously preserving information from the input.



$\hat{\mathbf{x}}$
$\hat{\mathbf{y}}$
$\mathbf{W}'_x,\mathbf{b}'_x$
$\mathbf{W}'_y,\mathbf{b}'_y$
$\mathbf{z}_x$
coupling
$\mathbf{z}_y$
$\mathbf{W}_x,\mathbf{b}_x$
$\mathbf{W}_y,\mathbf{b}_y$
Source (x)
Target (y)

FIGURE 5. **A CpAE is a pair of AEs where the hidden units (latent features) are coupled. The latent features, $\mathbf{z}_x$ and $\mathbf{z}_y$, are computed from the source and domain inputs, x and y, and the encoder parameters: $\mathbf{W}_x$, $\mathbf{b}_x$ and $\mathbf{W}_y$, $\mathbf{b}_y$. Additionally, source and domain reconstructions, $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, are computed using the latent features and decoder parameters: $\mathbf{W}'_x$, $\mathbf{b}'_x$ and $\mathbf{W}'_y$, $\mathbf{b}'_y$.**

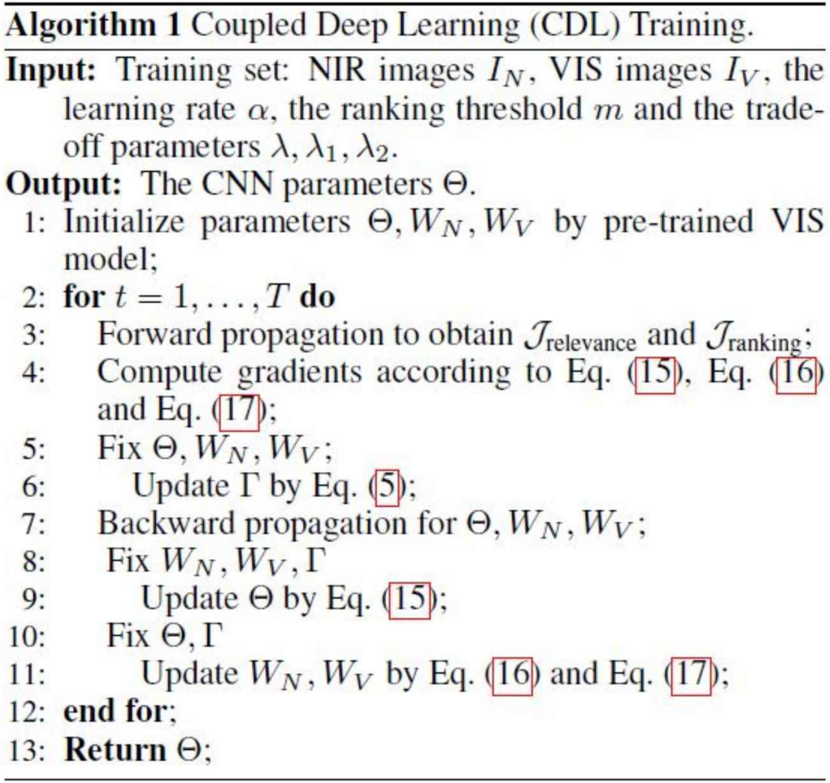


To Decoder/Classifier
$z_{x,2}$
coupling
$z_{y,2}$
$\mathbf{W}_{x,2},\mathbf{b}_{x,2}$
$\mathbf{W}_{y,2},\mathbf{b}_{y,2}$
$z_{x,1}$
coupling
$z_{y,1}$
$\mathbf{W}_{x,1},\mathbf{b}_{x,1}$
$\mathbf{W}_{y,1},\mathbf{b}_{y,1}$
Source (x)
Target (y)

FIGURE 6. **A stacked CpAE is a pair of stacked AEs with one (or more) coupled layers of hidden units. As shown, a subsequent CpAE is trained using the hidden units from the previous CpAE. For convenience, we have dropped the decoders.**

Wu et al (2017b)

- ✔ A coupled Deep Learning approach
- ✔ Transform HFR problem into homogeneous face matching problem by seeking a shared feature space
- ✔ Employ light CNN as the basic network
- ✔ To address the small-scale data for NIR-VIS training, firstly train a CNN on the large visible light face dataset and then fine-tune the NIR-VIS one on the pre-trained visible light face model
- ✔ Based on the basic network, develop a coupled deep learning (CDL) framework for NIR-VIS face recognition
- ✔ Combine the softmax term with relevance constraints and cross modal ranking term as the supervised signal



Algorithm 1 Coupled Deep Learning (CDL) Training.

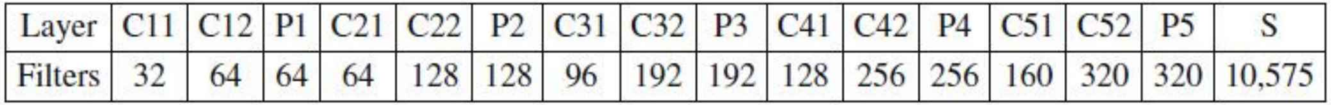
```
Input: Training set: NIR images $I_N$, VIS images $I_V$, the learning rate $\alpha$, the ranking threshold $m$ and the trade-off parameters $\lambda, \lambda_1, \lambda_2$.
Output: The CNN parameters $\Theta$.
Initialize parameters $\Theta, W_N, W_V$ by pre-trained VIS model;
for $t = 1, \ldots, T$ do
    Forward propagation to obtain $\mathcal{J}_{\text{relevance}}$ and $\mathcal{J}_{\text{ranking}}$;
    Compute gradients according to Eq. (15), Eq. (16) and Eq. (17);
    Fix $\Theta, W_N, W_V$;
        Update $\Gamma$ by Eq. (5);
    Backward propagation for $\Theta, W_N, W_V$;
     Fix $W_N, W_V, \Gamma$
        Update $\Theta$ by Eq. (15);
     Fix $\Theta, \Gamma$
        Update $W_N, W_V$ by Eq. (16) and Eq. (17);
end for;
Return $\Theta$;
```

Wu X, Song L, He R, Tan T (2017b) Coupled deep learning for heterogeneous face recognition. arXiv preprint arXiv:170402450

Saxena and Verbeek (2016)

✔ Explore different metric learning strategies to reduce discrepancies between different modalities

- NIR-VIS face recognition
- VIS-Sketch



Layer	C11	C12	P1	C21	C22	P2	C31	C32	P3	C41	C42	P4	C51	C52	P5	S
Filters	32	64	64	64	128	128	96	192	192	128	256	256	160	320	320	10,575

Fig. 2. CNN architecture: convolutions (C) use 3 × 3 filters and stride 1, max-pooling (P) act on 2 × 2 regions and use stride 2. The final soft-max classification layer is denoted as S.

Saxena S, Verbeek J (2016) Heterogeneous face recognition with cnns. In: Computer Vision–ECCV Workshops, Springer, pp 483–491

Kan et al (2016): multi-view deep network (MvDN)

- ✔ To eliminate the complex (maybe even highly nonlinear) view discrepancy for favorable cross-view recognition
- ✔ seeks for a non-linear discriminant and view-invariant representation shared between multiple views
- ✔ consists of two sub-networks
  - view-specific sub-network: attempt to remove view-specific variations
  - common sub-network: attempt to obtain common representation shared by all views

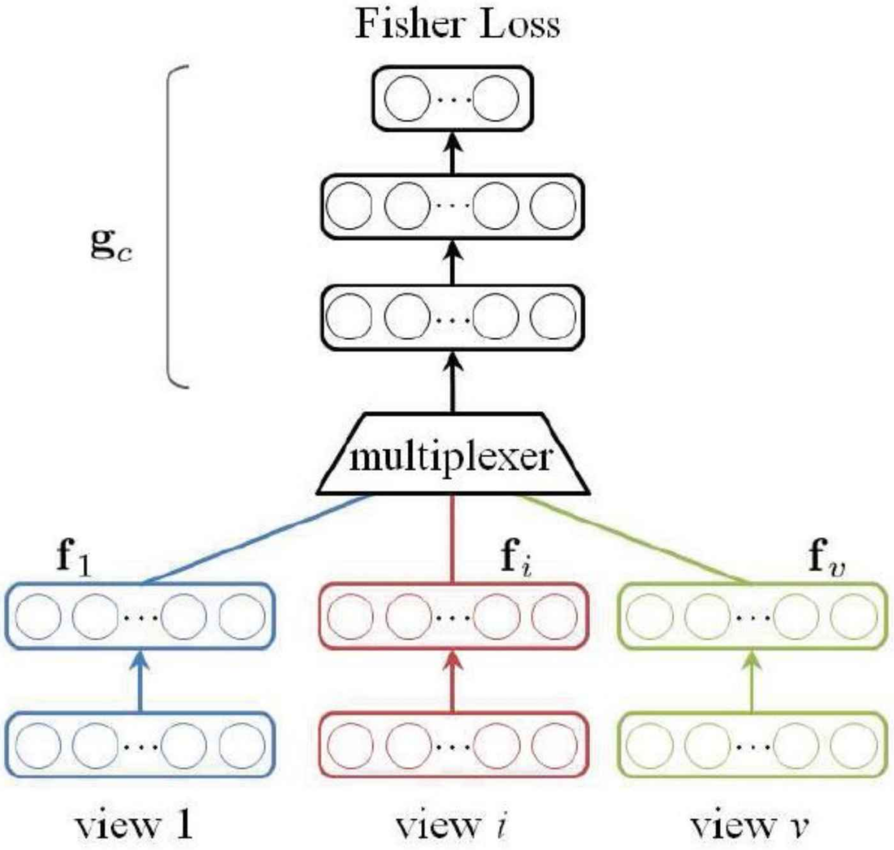
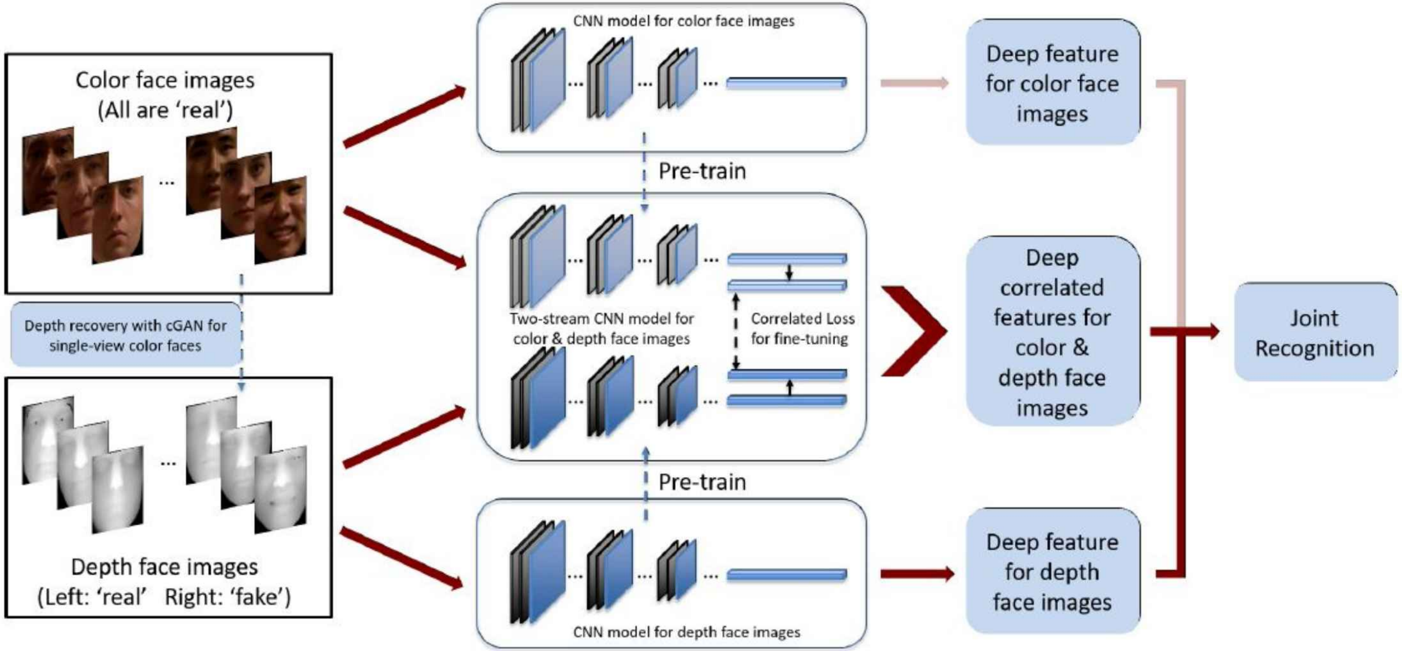


Figure 1. An overview of Multi-view Deep Network (MvDN). MvDN consists of two sub-networks, the view-specific sub-network $\mathbf{f}_i|_{i=1}^{v}$ and the common sub-network $\mathbf{g}_c$, along with a discriminant Fisher objective.

Kan M, Shan S, Chen X (2016) Multi-view deep network for cross-view classification. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition, pp 4847–4855

Zhang et al (2017c) -- see GAN+CNN

✔ Combine the generative capacity of conditional GAN and the discriminative feature extraction of DCNN for cross-modality learning



Color face images
(All are 'real')
Depth recovery with cGAN for single-view color faces
Depth face images
(Left: 'real' Right: 'fake')
CNN model for color face images
Pre-train
Two-stream CNN model for color & depth face images
Correlated Loss for fine-tuning
Pre-train
CNN model for depth face images
Deep feature for color face images
Deep correlated features for color & depth face images
Deep feature for depth face images
Joint Recognition

Figure 1: Overview of the proposed CNN models for heterogeneous face recognition. Note that (1) depth recovery is conducted only for testing; (2) the final joint recognition may or may not include color based matching, depending on the specific experiment protocol.

Zhang W, Shu Z, Samaras D, Chen L (2017c) Improving heterogeneous face recognition with conditional adversarial networks. arXiv preprint arXiv:170902848

Cross-moda l CNN Model

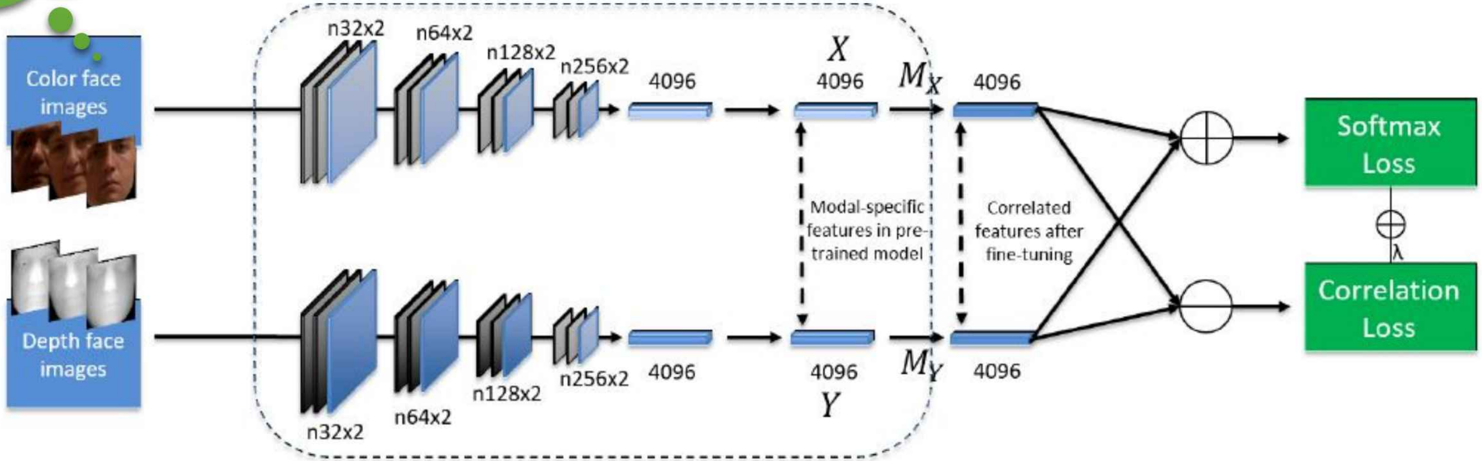


Figure 3: Training procedure of the cross-modal CNN model. Models in the dashed box are pre-trained using 2D and 2.5D face images individually.

- Once a pair of unimodal models for both views (depth and color) are trained, the modal-specific representations, {X,Y}, can be obtained after the last fully connected layers
- a joint supervision is required to enforce both correlation and distinctiveness simultaneously

Song et al (2017)

- ✔ An adversarial discriminative feature learning framework
- ✔ close the gap between sensing patterns of different face modalities on both raw-pixel space and compact feature space
- ✔ integrates cross-spectral face hallucination and discriminative feature learning into an end-to-end adversarial network

- In the pixel space:
  - ○ use generative adversarial networks to perform cross-spectral face hallucination. An elaborate two-path model is introduced to alleviate the lack of paired images, which gives consideration to both global structures and local textures.
- In the feature space:
  - ○ an adversarial loss and a high-order variance discrepancy loss are employed to measure the global and local discrepancy between two heterogeneous distributions respectively.
  - ○ These two losses enhance domain-invariant feature learning and modality independent noise removing

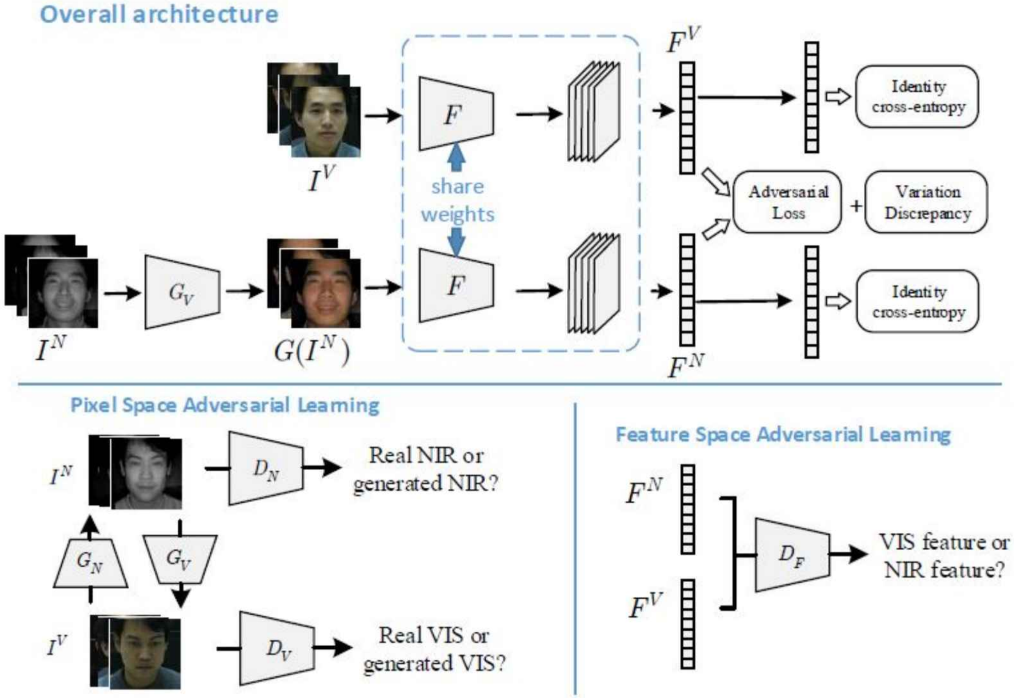


Figure 1: The proposed adversarial discriminative HFR framework. Adversarial learning is employed on both raw-pixel space and compact feature space.

Song L, Zhang M, Wu X, He R (2017) Adversarial discriminative heterogeneous face recognition. arXiv preprint arXiv:170903675

- There are some typical Heterogenous Face Recognition
  - □ Still-to-Video
  - □ NIR/IR-VIS
  - □ Photo-Sketch
  - □ others

NIR/IR-VIS

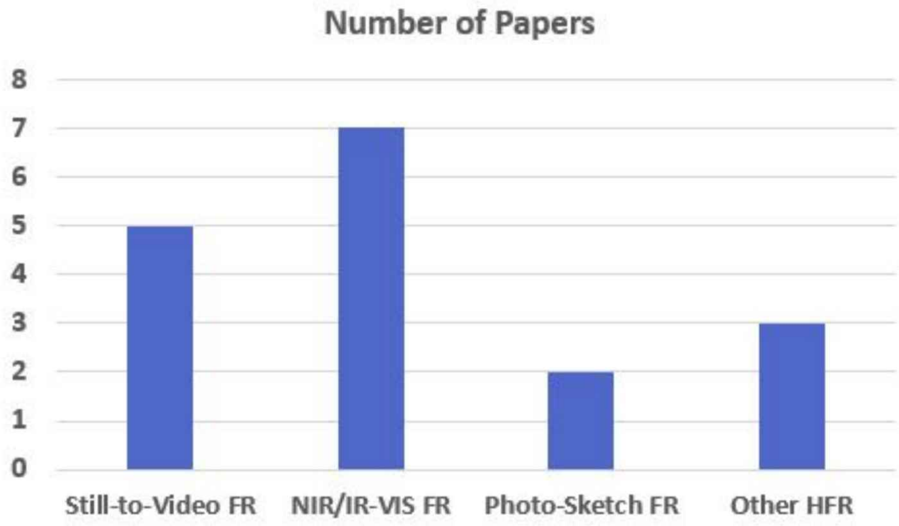


Photo-Sketch

❖ *Still-to-Video Face Recognition (S2V)*

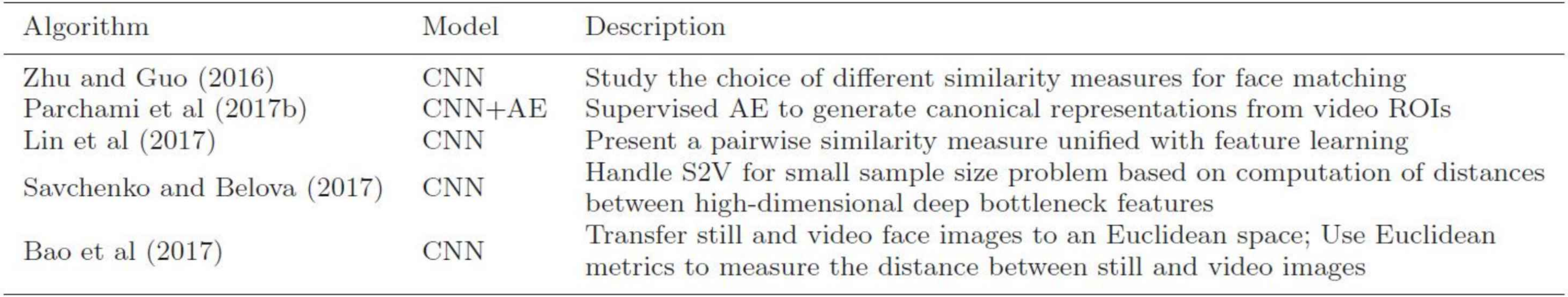
- S2V face recognition has real-world applications
- Usually, the gallery set has higher resolution still images
- while the probe is video clips with lower resolutions

 Zhu et al (2015)

- ✔ addressed the S2V face recognition problem as heterogeneous face matching
- ✔ used a domain adaptation method for S2V

- Some deep methods have been proposed to bridge the gap between these two modalities

Table 17 Overview of deep learning methods for S2V face recognition



Algorithm	Model	Description
Zhu and Guo (2016)	CNN	Study the choice of different similarity measures for face matching
Parchami et al (2017b)	CNN+AE	Supervised AE to generate canonical representations from video ROIs
Lin et al (2017)	CNN	Present a pairwise similarity measure unified with feature learning
Savchenko and Belova (2017)	CNN	Handle S2V for small sample size problem based on computation of distances between high-dimensional deep bottleneck features
Bao et al (2017)	CNN	Transfer still and video face images to an Euclidean space; Use Euclidean metrics to measure the distance between still and video images

CFR-CNN (Parchami et al, 2017b)

- ✔ an efficient Canonical Face Representation CNN for S2V face recognition
- ✔ uses a supervised autoencoder network to generate canonical face representations from video regions of interest

✔consists of two major components:

- o autoencoder : to learn discriminant face embeddings, and to reconstruct a high-quality canonical ROIs
- o classification networks: matches the face embeddings for a pair of reference still and probe video ROIs

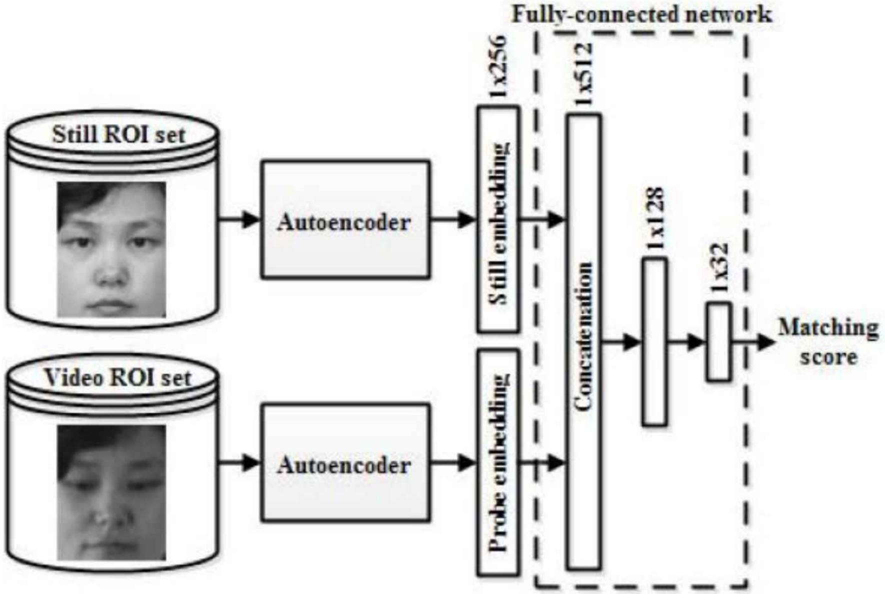


Figure 3: Block diagram of the proposed CFR-CNN.

Parchami M, Bashbaghi S, Granger E, Sayed S (2017b) Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In: Advanced Video and Signal Based Surveillance, Intl. Conf. on, IEEE, pp 1–6

- autoencoder
  - ✔ learn discriminant face embeddings
  - ✔ reconstruct a high-quality canonical ROIs
    - frontal, well-illuminated, less blurred faces with neutral expression

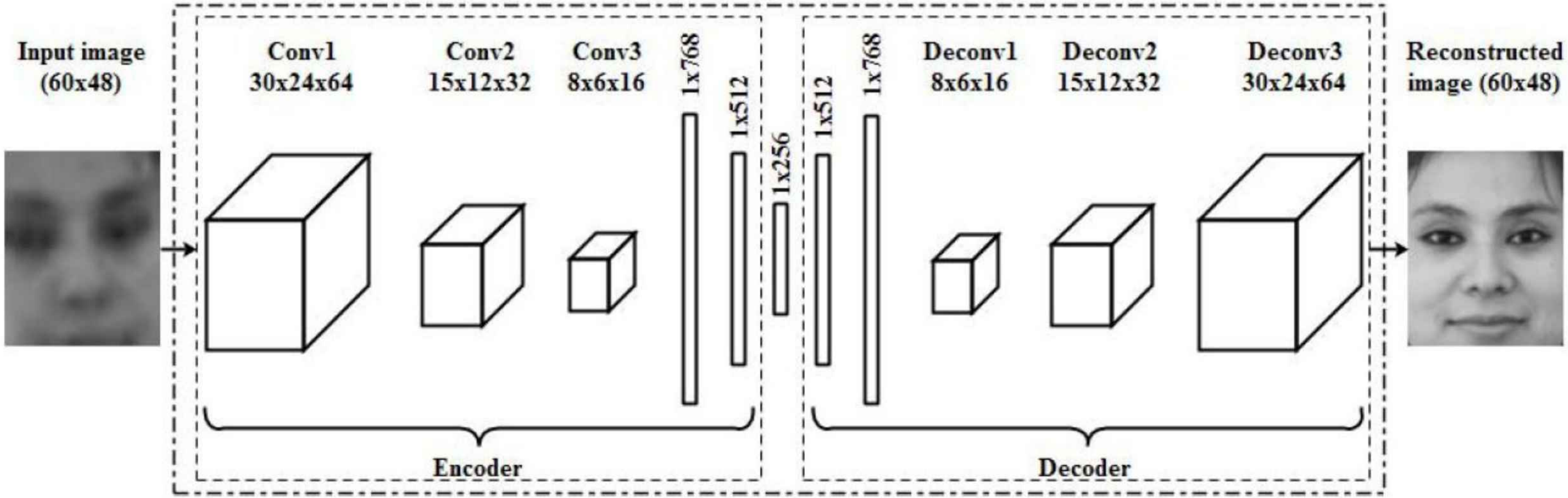


Figure 1: Block diagram of the proposed autoencoder network in the CFR-CNN.

Lin et al (2017)

- ✔ presented a pairwise similarity measure and unified it with feature representation learning via DCNN
- ✔ This model can be used to handle S2V problem

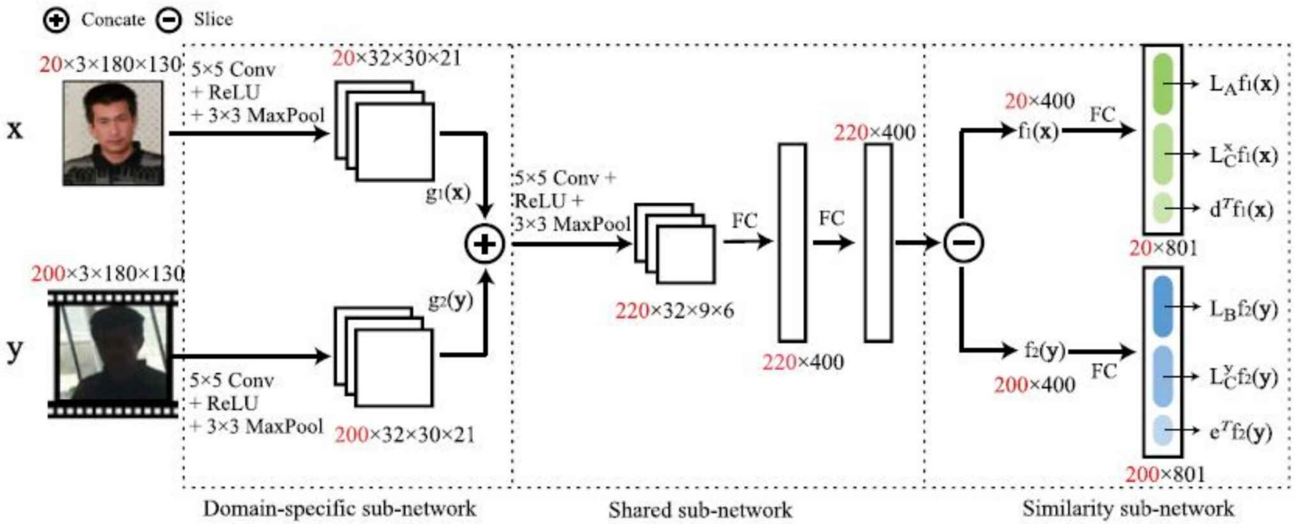


Fig. 3. Deep architecture of our similarity model. This architecture is comprised of three parts: domain-specific sub-network, shared sub-network and similarity sub-network. The first two parts extract feature representations from samples of different domains, which are built upon a number of convolutional layers, max-pooling operations and fully-connected layers. The similarity sub-network includes two structured fully-connected layers that incorporate the similarity components in Eqn. (3).

Lin L, Wang G, Zuo W, Feng X, Zhang L (2017) Cross-domain visual matching via generalized similarity measure and feature learning. trans on pattern analysis and machine intelligence 39(6):1089–1102

Bao et al (2017)

- ✔ transferred still and video face images to an Euclidean space
- ✔ adopted Euclidean metrics to measure the distance between still and video images
- ✔ loss function are designed as a regression one to minimize the intra-class variations while maximize the inter-class variations

Bao T, Ding C, Karmoshi S, Zhu M (2017) Video-based face recognition via convolutional neural networks. In: Second Intl. Workshop on Pattern Recognition, Intl. Society for Optics and Photonics, vol 10443, p 104430I

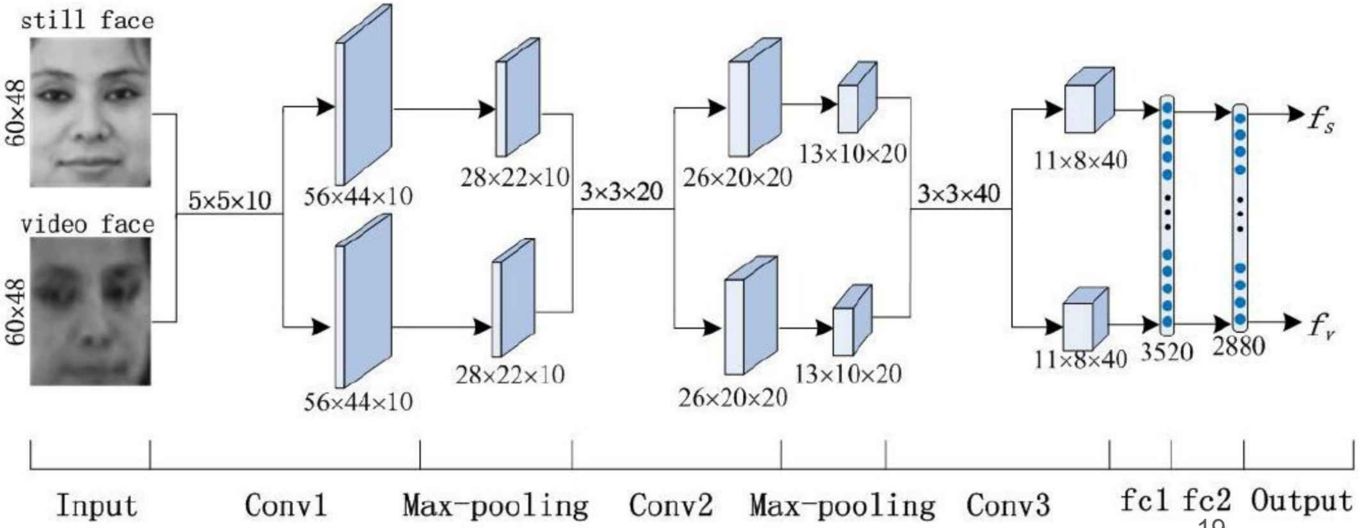


Figure 1. The deep architecture of CNN for feature extraction.

Savchenko and Belova (2017)

✔ addressed S2V face recognition for small sample size problem

✔ using a statistical recognition method which casts S2V into Maximum A Posteriori estimation

❖ *NIR/IR-VIS Face Recognition*

- Infrared spectra have different regions:
  - reflection dominated region
    - ✔ near infrared (NIR)
    - ✔ shortwave infrared (SWIR) bands;
  - emission dominated thermal region
    - ✔ midwave infrared (MWIR)
    - ✔ longwave infra-red (LWIR) bands
- The main advantage of thermal imaging:
  - the acquisition in low light conditions where the visible light cameras cannot work
- Matching thermal face images to the visible is quite challenging, and has made limited progress.

- In recent years, there has been some growing interest in the infrared spectrum for face recognition
- Deep learning techniques have been applied

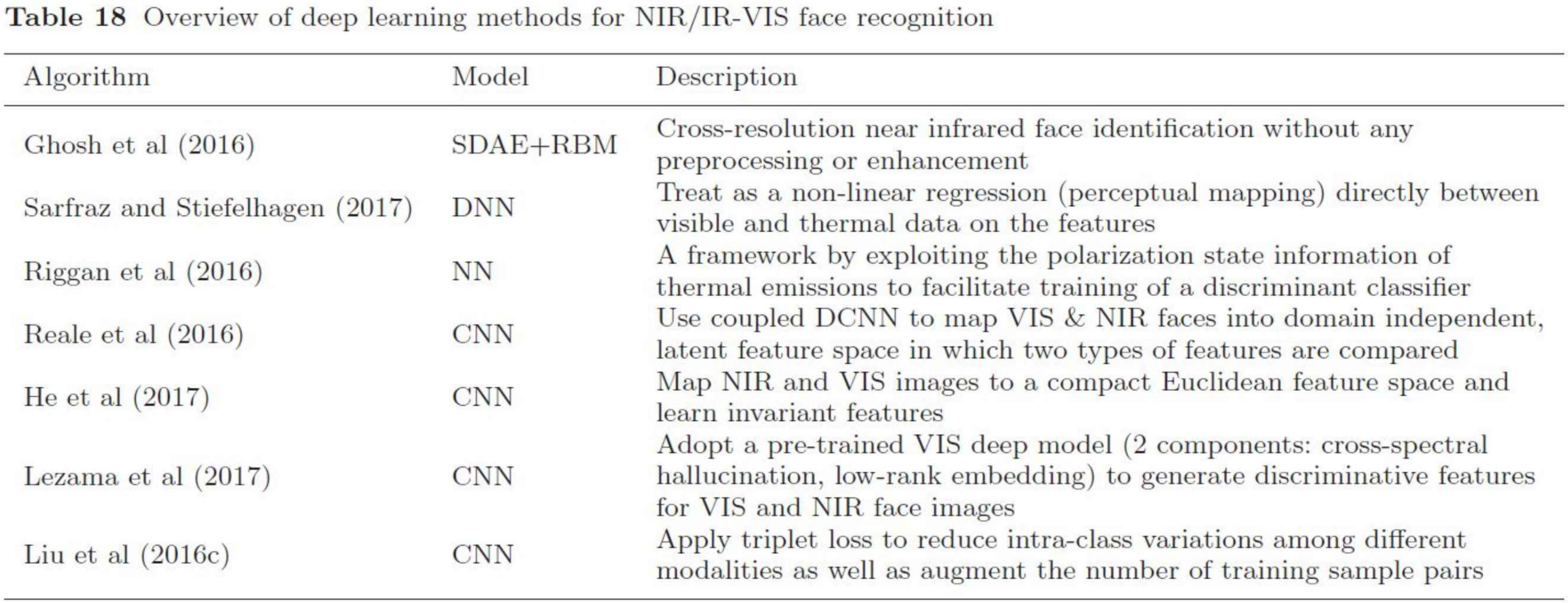


Table 18 Overview of deep learning methods for NIR/IR-VIS face recognition

Algorithm	Model	Description
Ghosh et al (2016)	SDAE+RBM	Cross-resolution near infrared face identification without any preprocessing or enhancement
Sarfraz and Stiefelhagen (2017)	DNN	Treat as a non-linear regression (perceptual mapping) directly between visible and thermal data on the features
Riggan et al (2016)	NN	A framework by exploiting the polarization state information of thermal emissions to facilitate training of a discriminant classifier
Reale et al (2016)	CNN	Use coupled DCNN to map VIS & NIR faces into domain independent, latent feature space in which two types of features are compared
He et al (2017)	CNN	Map NIR and VIS images to a compact Euclidean feature space and learn invariant features
Lezama et al (2017)	CNN	Adopt a pre-trained VIS deep model (2 components: cross-spectral hallucination, low-rank embedding) to generate discriminative features for VIS and NIR face images
Liu et al (2016c)	CNN	Apply triplet loss to reduce intra-class variations among different modalities as well as augment the number of training sample pairs

▪ Thermal-to-Visible face matching

- One type of cross-modal face recognition
- Identify a thermal infrared (IR) face image given a gallery of visible light face images
- The main idea is to:
  - exploit structural similarities between visible (VIS) and thermal infrared (IR) facial imagery

 Riggan et al (2016)

✔ exploited the polarization state information of thermal emissions for polarimetric thermal-to-visible face recognition with a polarimetric thermal imaging technique

DPM (Sarfraz and Stiefelhagen, 2017)

- ✔ Deep Perceptual Mapping directly learned a mapping from visible features to thermal or polarimetric features, or vice versa.
- ✔ Deep neural networks can learn, to some extent, the non-linear mapping by adjusting the projection coefficients in an iterative manner over the training set.

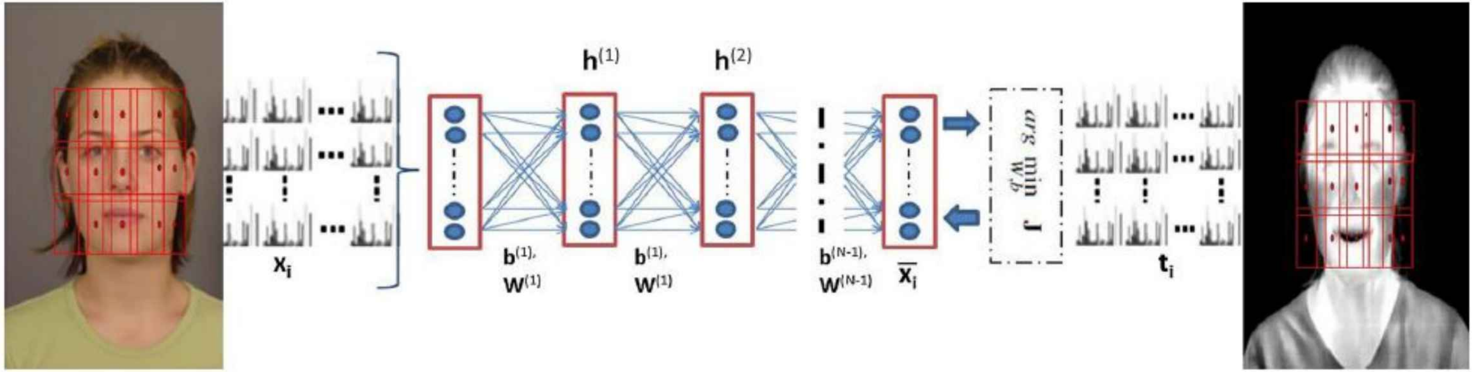


Fig. 1 Deep perceptual mapping (DPM): densely computed features from the visible domain are mapped through the learned DPM network to the corresponding thermal domain

Sarfraz MS, Stiefelhagen R (2017) Deep perceptual mapping for cross-modal face recognition. Intl Journal of Computer Vision 122(3):426–438

▪ NIR-VIS face recognition

 Reale et al (2016)

- ✔ used coupled DCNNs to map VIS and NIR faces into a domain independent, latent feature space in which two types of features can be compared directly

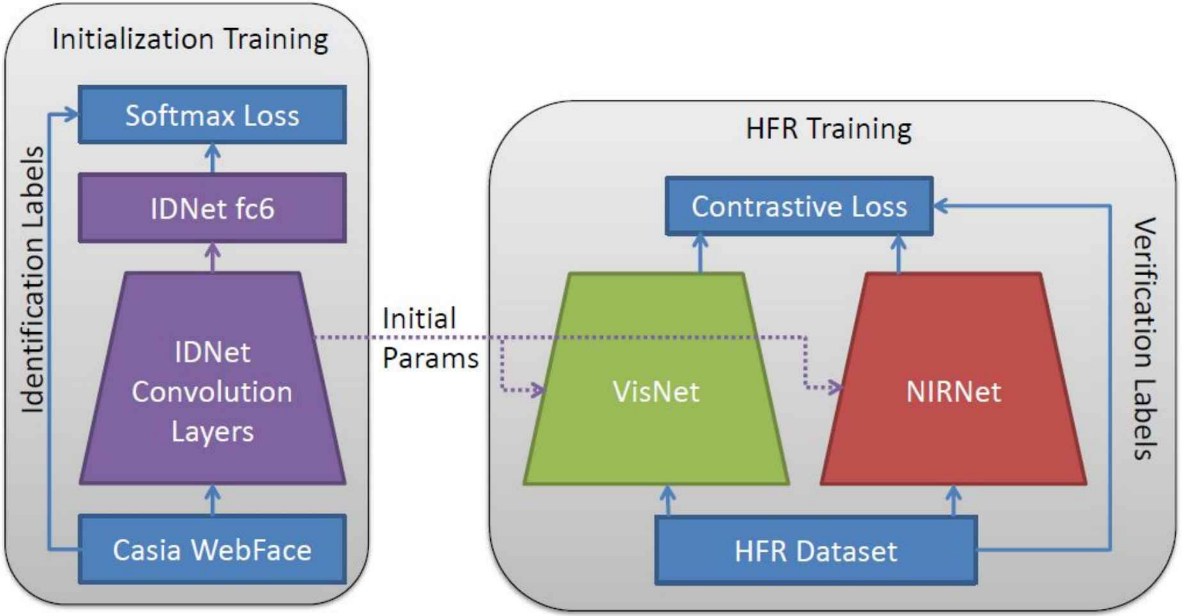


Figure 3: Network Diagram

- ✔ The network structure is from the GoogLeNet family of networks (i.e. deep with small convolutional filters)
- ✔ train two networks for cross-modal verification:
  - ○ initialize these networks as copies of IDNet, with the exclusion of the fully connected softmax classifier

Liu et al (2016c)

- ✔ applied the triplet loss to:
  - ○ reduce intra-class variations among different modalities
  - ○ augment the number of training sample pairs

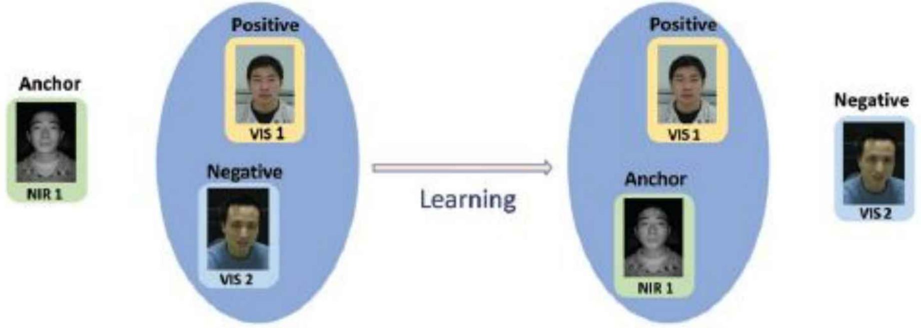


Figure 1. NIR-VIS triplet formation. After learning, the distance between the samples from the same ID is minimized while the difference across domains tends to be not the leading factor. [Best viewed in color]

Liu X, Song L, Wu X, Tan T (2016c) Transferring deep representation for nir-vis heterogeneous face recognition. In: Biometrics, Intl. Conf. on, IEEE, pp 1–8

✔ With the constraint of the triplet loss, discriminative features can be learned to differentiate different identities no matter which modality they belong to, NIR or VIS

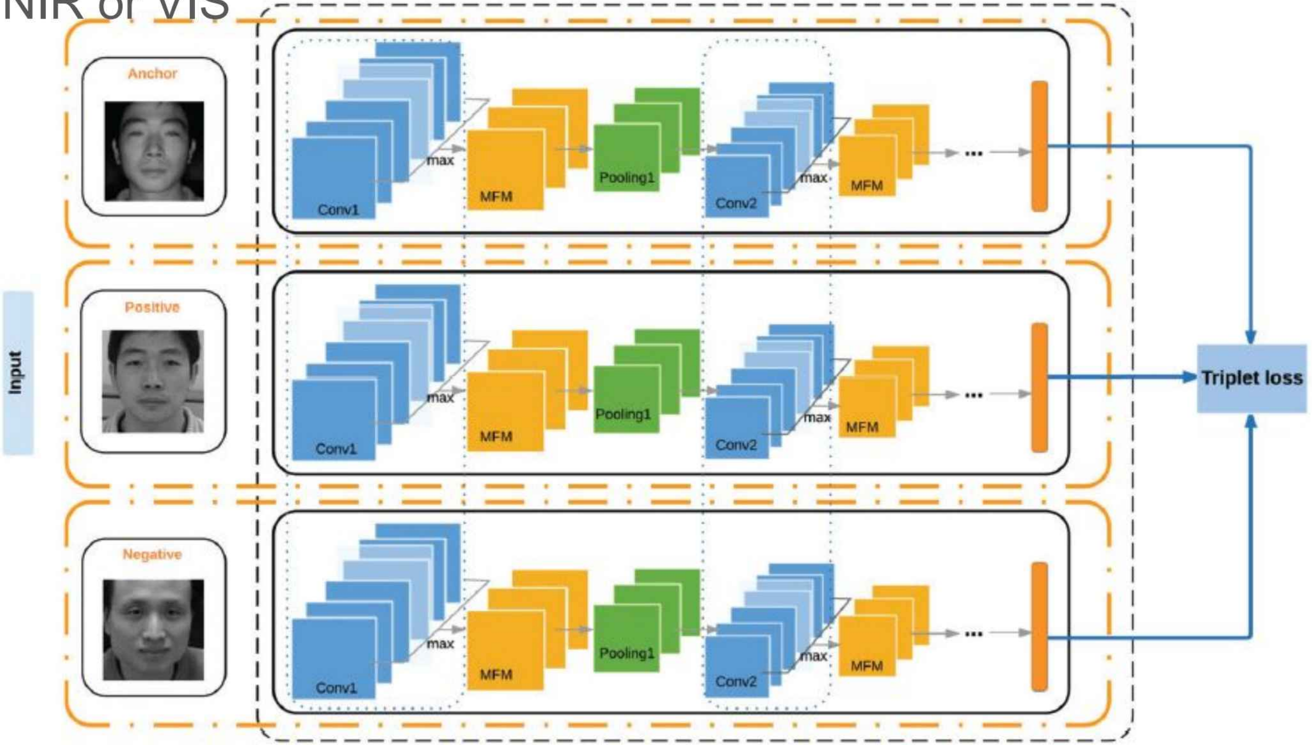


Figure 2. The proposed CNN framework. The inputs of CNN are the prepared triplets, and the three channels share the same parameters. After feature extraction in the final fully connected layers, the high-level features of the three layers are input to the triplet loss layer which bridges the gap of NIR and VIS domains.

Lezama et al (2017)

Lezama J, Qiu Q, Sapiro G (2017) Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In: Conf. on Computer Vision and Pattern Recognition, IEEE, pp 6807–6816

- ✔ Extend a DNN model pre-trained on VIS face images to the NIR spectrum
- ✔ Add two extra steps:
  - ○ VIS hallucination CNN:
    - ▪ preprocess the NIR image using a CNN that performs a cross-spectral conversion of the NIR image into the VIS spectrum
    - ▪ using the hallucinated VIS image as input to the feature extraction DNN, instead of the raw NIR, produces a significant gain in the recognition performance.
  - ○ use low-rank embedding at the output layer:
    - ▪ Can produces deep features for VIS and NIR images in a common space
    - ▪ A geometrically motivated transformation is learned to restore a within-class low-rank structure, and meanwhile introduce a maximally separated inter-class structure.

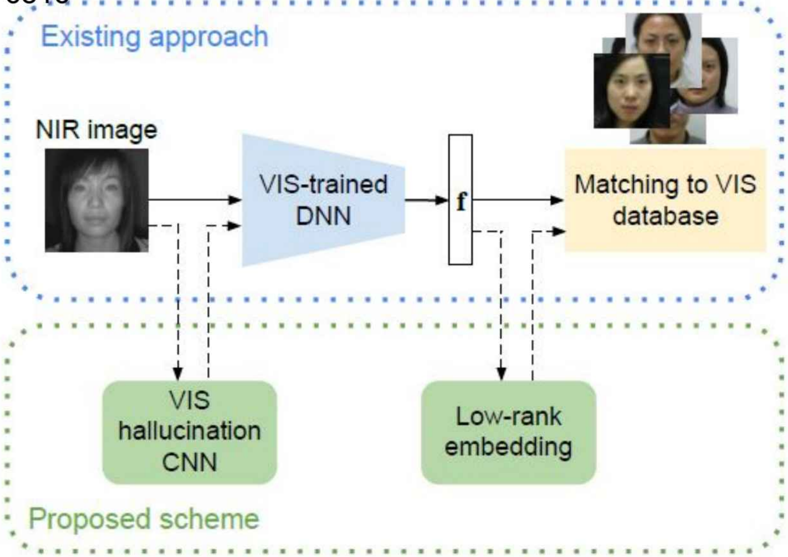


Figure 1. Diagram of the proposed approach. A simple NIR-VIS face recognition system consists in using a Deep Neural Network (DNN) trained only on VIS images to extract a feature vector **f** from a NIR image and use it for matching to a VIS database. We propose two modifications to this basic system. First, we modify the input by hallucinating a VIS image from the NIR sample. Secondly, we apply a low-rank embedding of the DNN features at the output. Each of this modifications produces important improvements in the recognition performance, and an even greater one when applied together.

He et al (2017)

He R, Wu X, Sun Z, Tan T (2017) Wasserstein cnn: Learning invariant features for nir-vis face recognition. arXiv preprint arXiv:170802412

✔ mapped both NIR and VIS images to a compact Euclidean feature space and learned invariant features

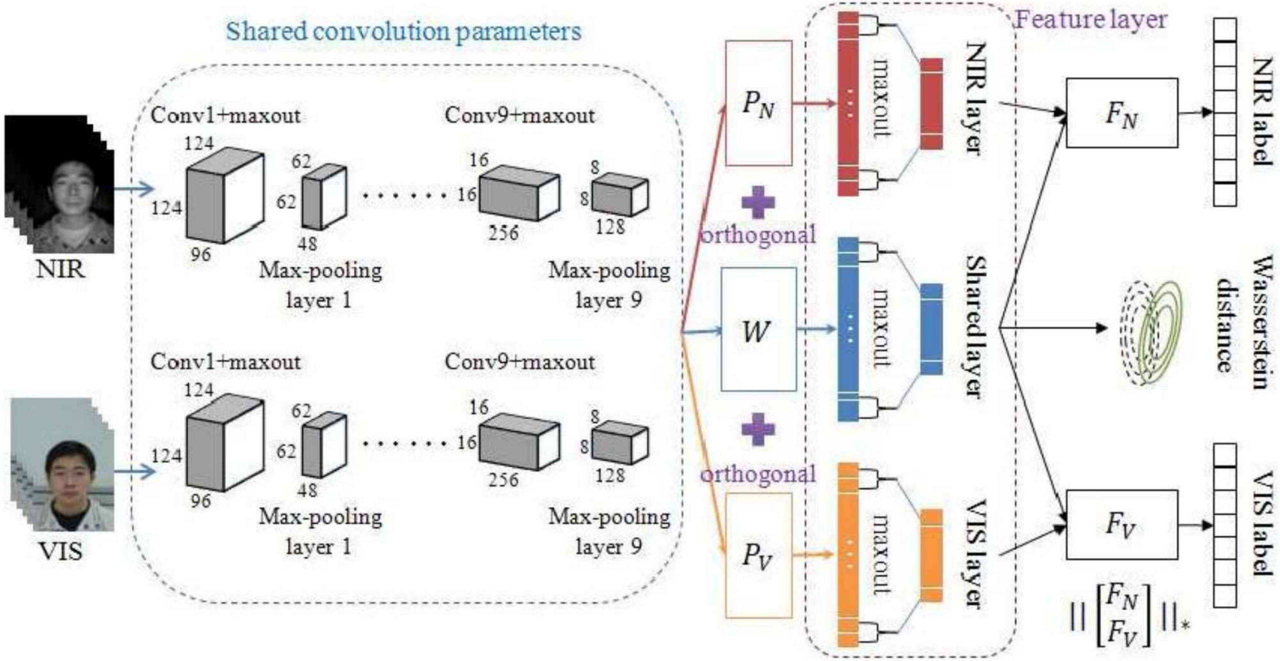
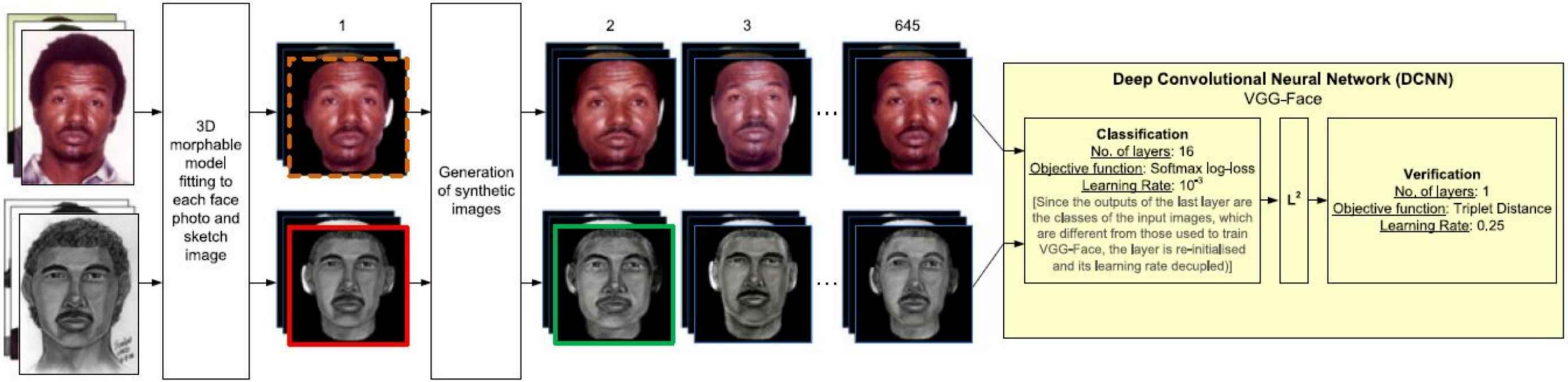


Fig. 1. An illustration of our proposed Wasserstein CNN architecture. The Wasserstein distance is used to measure the difference between NIR and VIS distributions in the modality invariant subspace (spanned by matrix W). At the testing time, both NIR and VIS features are exacted from the shared layer of one single neural network and compared in cosine distance.

Galea C, Farrugia RA (2017) Forensic face photo-sketch recognition using a deep learning-based architecture. IEEE Signal Processing Letters 24(11):1586–1590

❖ *Photo-Sketch Face Recognition*

- Galea and Farrugia (2017)
  - ✔ A model which is pre-trained for face photo recognition and tuned for photo-sketch matching by applying the transfer learning


1
2
3
645
3D morphable model fitting to each face photo and sketch image
Generation of synthetic images
...
...
Deep Convolutional Neural Network (DCNN)
VGG-Face
Classification
No. of layers: 16
Objective function: Softmax log-loss
Learning Rate: 10^{-3}
L^2
Verification
No. of layers: 1
Objective function: Triplet Distance
Learning Rate: 0.25

[Since the outputs of the last layer are the classes of the input images, which are different from those used to train VGG-Face, the layer is re-initialised and its learning rate decupled)]

Fig. 1. Proposed architecture, where synthetic images are created and used to train the DCNN in [9] via transfer learning. The first and second rows contain original and synthesised photos and sketches, respectively, of a subject in the PRIP-HDC forensic sketch database [12]. Column '1' contains images fitted with a 3-D Morphable model, and "2" to "645" are synthesised versions of "1". The synthetic sketch of variation "2" (represented with a green border) has a more rounded appearance than the original sketch (red border) and bears a subjectively better similarity to the corresponding original photo (dashed orange border). As shown in the yellow box, the DCNN is first trained for classification and then tuned for verification using triplet embedding.

Lin L, Wang G, Zuo W, Feng X, Zhang L (2017) Cross-domain visual matching via generalized similarity measure and feature learning. trans on pattern analysis and machine intelligence 39(6):1089–1102

- Lin et al (2017)
  - ✔ present a pairwise similarity measure
  - ✔ unify it with feature representation learning via deep convolutional neural networks,
  - ✔ used for photo-sketch face matching

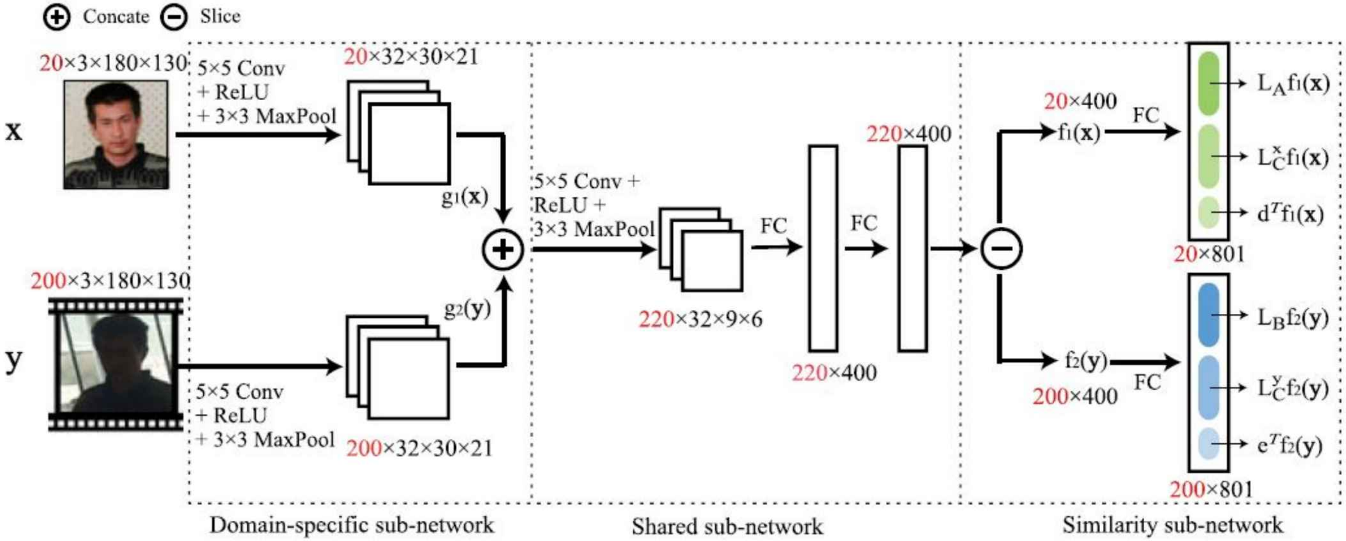


Fig. 3. Deep architecture of our similarity model. This architecture is comprised of three parts: domain-specific sub-network, shared sub-network and similarity sub-network. The first two parts extract feature representations from samples of different domains, which are built upon a number of convolutional layers, max-pooling operations and fully-connected layers. The similarity sub-network includes two structured fully-connected layers that incorporate the similarity components in Eqn. (3).

❖ *Other Heterogeneous Face Recognition*

- Simon et al (2016):
  - ✔ applied deep CNNs to the tri-modal RGB-D-T based facial recognition problem

The result shows that:

✔ in most cases, using such 3 modalities provides a better identification performance than an isolated or bimodal approach.

Sim´on MO, Corneanu C, Nasrollahi K, Nikisins O, Escalera S, Sun Y, Li H, Sun Z, Moeslund TB, Greitans M (2016) Improved rgb-dt based face recognition. Iet Biometrics 5(4):297–303

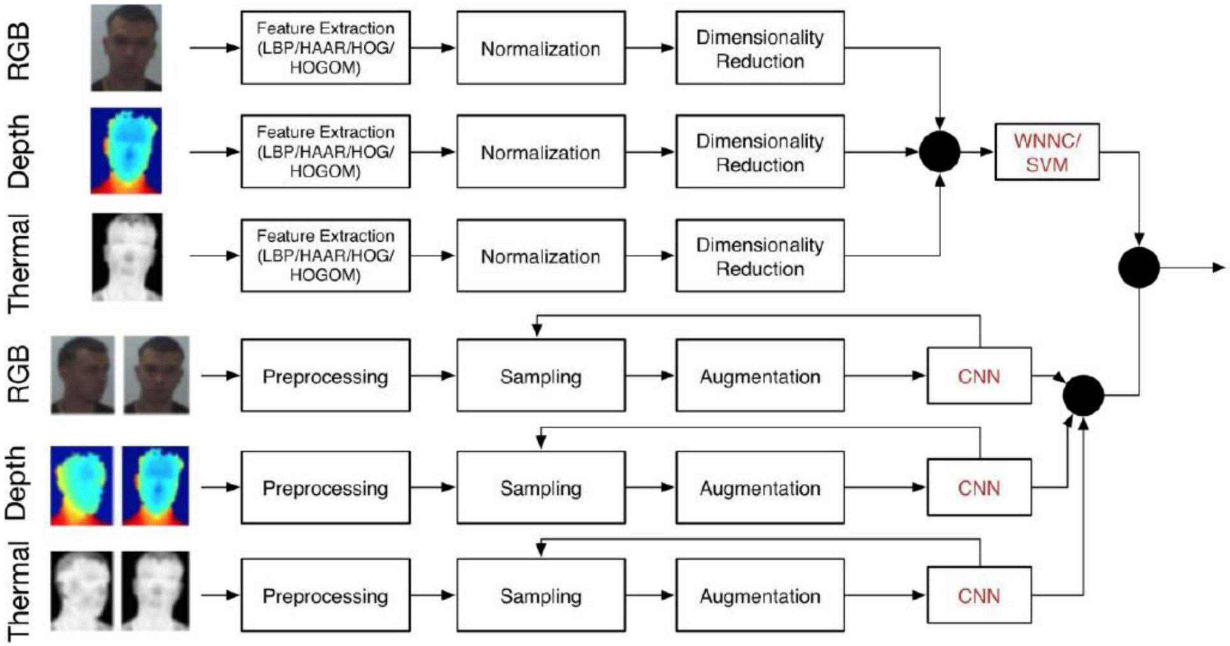
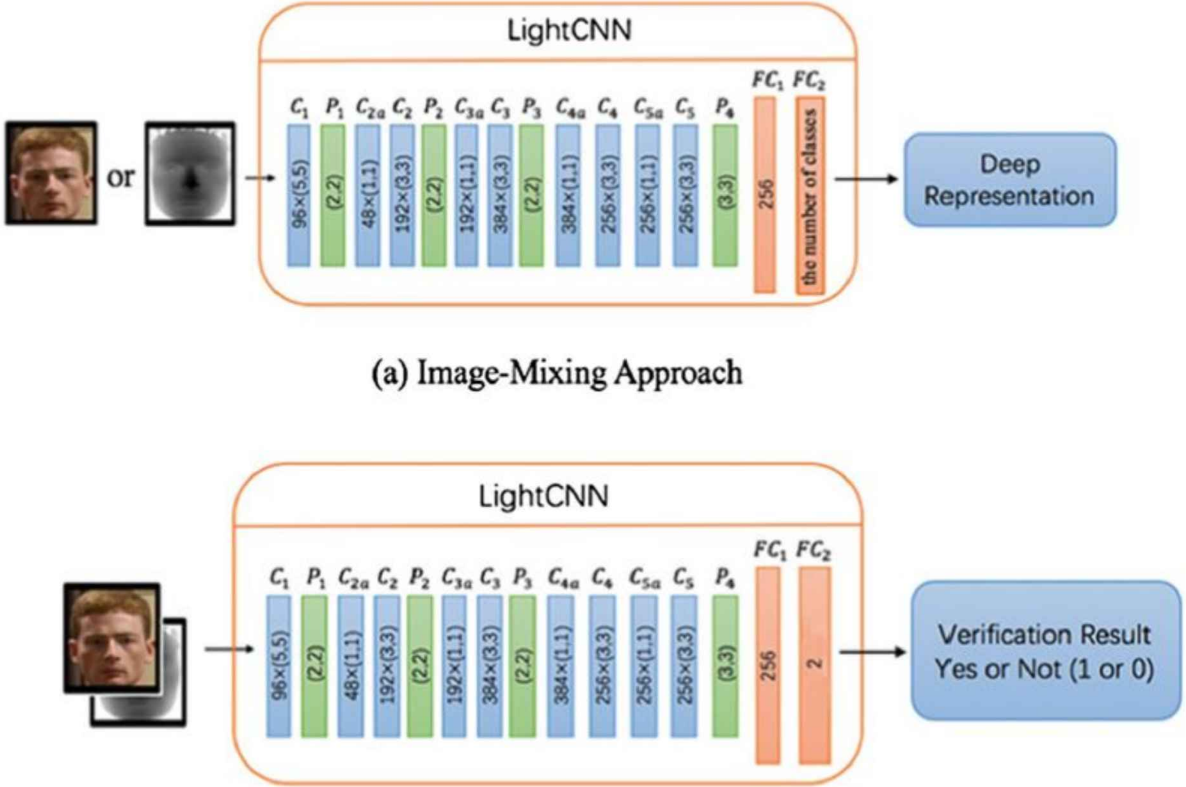


Figure 1: The block diagram of the proposed system. RGB, Depth and Thermal captures of the face are used for training modality specific CNNs for deciding if two samples are from the same person or not. The results are fused with a HOGOM trained WNNC and SVM.

- Liu et al (2017a)
  - ✔ two deep CNNs based approach for Depth-to-RGB face recognition
  - ✔ employ lightCNN as the baseline
  - ✔ In order to enable LightCNN to extract discriminative features from both RGB and Depth face images, the image-mixing approach simply mixes the RGB and Depth face images to form a single training set. This training set is then used to finetune the original LightCNN
  - ✔ Once the LightCNN is finetuned by the mixed training set of RGB and Depth face images, it can be used to extract features from RGB or Depth face images by exporting the output of the *FC*1 layer.
  - ✔ Based on the extracted features, the Cosine distance metric [21] is used to measure the similarity between a pair of RGB and Depth face images.

Liu H, He F, Zhao Q, Fei X (2017a) Matching depth to rgb for boosting face verification. In: Chinese Conf. on Biometric Recognition, Springer, pp 127–134



(a) Image-Mixing Approach

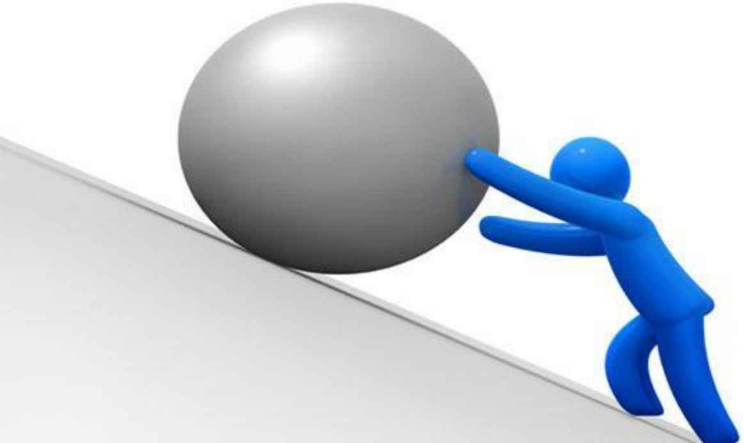
(b) Image-Fusion Approach

Fig. 2. Matching Depth to RGB face images, (a) is Image-Mixing Approach, (b) is Image-Fusion Approach. Note that RGB images are converted to gray images.

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



- Open-Set vs. Closed-Set Face Recognition

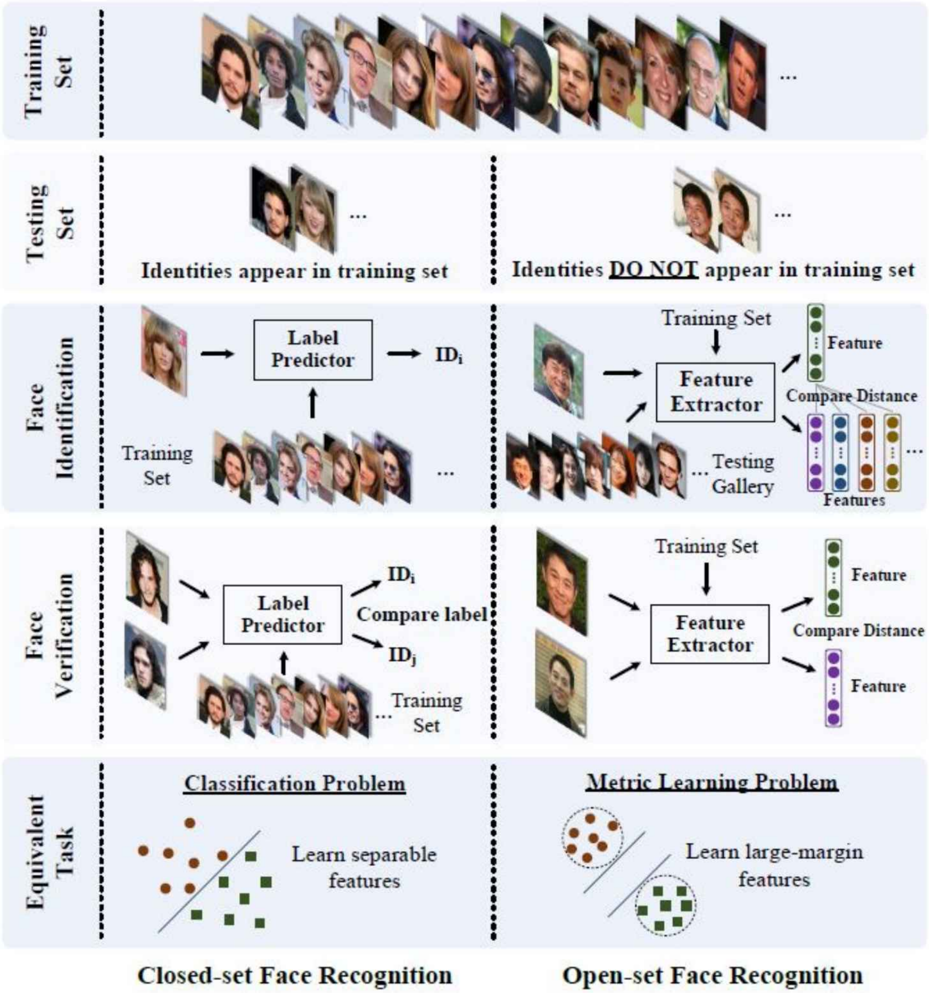


Figure 1: Comparison of open-set and closed-set face recognition.

- Gunther et al (2017)
  - ✔ evaluated the challenges for unconstrained open-set face recognition
  - ✔ Although face verification or closed-set face identification have surpassed human capabilities on some datasets, open-set identification is much more complex as it needs to reject both unknown identities and false accepts from the face detector
  - ✔ open-set face recognition is currently weak and requires much more attention

- Vareto et al (2017)
  - ✔ combined hashing functions and classification methods to estimate when probe samples are known (i.e., belong to the gallery set)
  - ✔ They did experiments with partial least squares and neural networks
  - ✔ and showed how response value histograms tend to behave for known and unknown individuals whenever they test a probe

- Wang et al (2017b) ---see Video
  - ✔ built a DCNN framework with a triplet supervisory signal
  - ✔ to identify few suspects from the crowd in real time for public video surveillance

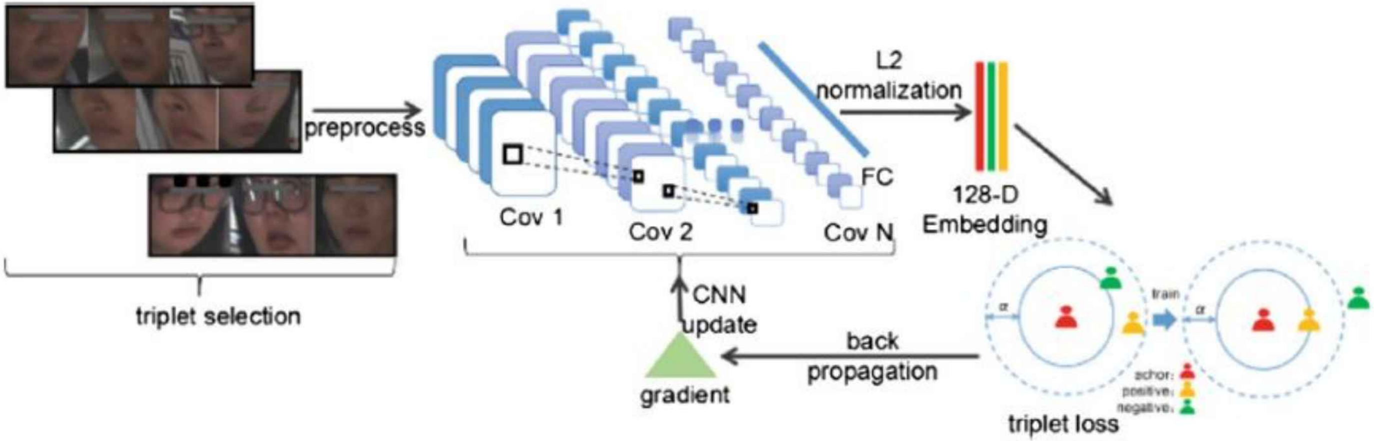


Fig. 2. End to end deep embedding training with triplet loss

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