Improving Heterogeneous Face Recognition with Conditional Adversarial Networks

Wuming Zhang[1]
wuming.zhang@ec-lyon.fr
Zhixin Shu[2]
zhshu@cs.stonybrook.edu
Dimitris Samaras[2]
samaras@cs.stonybrook.edu
Liming Chen[1]
liming.chen@ec-lyon.fr

[1] Laboratory LIRIS
Ecole Centrale de Lyon
Ecully, France

[2] Computer Vision Lab
Stony Brook University
Stony Brook, NY, USA

Abstract

Heterogeneous face recognition between color image and depth image is a much desired capacity for real world applications where shape information is looked upon as merely involved in gallery. In this paper, we propose a cross-modal deep learning method as an effective and efficient workaround for this challenge. Specifically, we begin with learning two convolutional neural networks (CNNs) to extract 2D and 2.5D face features individually. Once trained, they can serve as pre-trained models for another two-way CNN which explores the correlated part between color and depth for heterogeneous matching. Compared with most conventional cross-modal approaches, our method additionally conducts accurate depth image reconstruction from single color image with Conditional Generative Adversarial Nets (cGAN), and further enhances the recognition performance by fusing multi-modal matching results. Through both qualitative and quantitative experiments on benchmark FRGC 2D/3D face database, we demonstrate that the proposed pipeline outperforms state-of-the-art performance on heterogeneous face recognition and ensures a drastically efficient on-line stage.

1 Introduction

For decades, face recognition (FR) from color images has achieved substantial progress and forms part of an ever-growing number of real world applications, such as video surveillance, people tagging and virtual/augmented reality systems [3, 32, 40]. With the increasing demand for recognition accuracy under unconstrained conditions, the weak points of 2D based FR methods become apparent: as an imaging-based representation, color image is quite sensitive to numerous external factors, such as lighting variations and makeup patterns. Therefore, 3D based FR techniques [5, 7, 8] have recently emerged as a remedy because they take into consideration the intrinsic shape information of faces which is more robust while dealing with these nuisance factors. Moreover, the complementary strengths of color and depth data allow them to jointly work and gain further improvement.

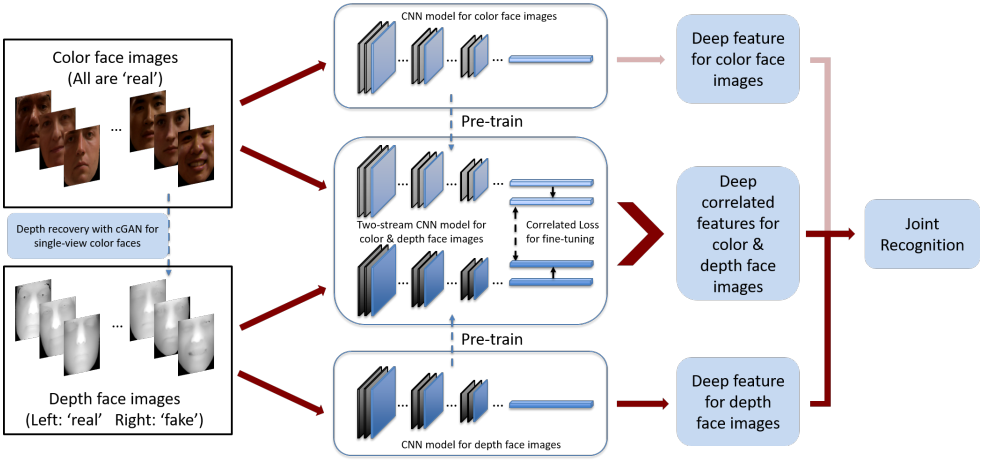


Figure 1: Overview of the proposed CNN models for heterogeneous face recognition. Note that (1) depth recovery is conducted only for testing; (2) the final joint recognition may or may not include color based matching, depending on the specific experiment protocol.

However, depth data is not always accessible in real-life conditions due to its special requirements for optical instruments and acquisition environment. Likewise, other challenges remain as well, including the real-time registration and preprocessing for depth images. An important question then naturally arises: can we design a recognition pipeline where depth images are only registered in gallery while still providing significant information for the identification of unseen color images? To cope with this problem, heterogeneous face recognition (HFR) [13, 33, 41] has been proposed as a reasonable workaround. As a worthwhile trade-off between purely 2D and 3D based method, HFR adopts both color and depth data for training and gallery set while the online probe set will simply contains color images. Under this mechanism, a HFR framework can take full advantage of both color and depth information at the training stage to reveal the correlation between them. Once learned, this cross-modal correlation makes it possible to conduct heterogeneous matching between preloaded depth images in gallery and color images digitally captured in real time.

Beyond the above-mentioned mechanism, in this paper we take a further look at our constraint on the use of depth image. Note that all difficulties, which hinder us from availing ourselves of depth information in probe set, come from the acquisition and registration of 3D data. Intuitively, these problems can be immediately solved if we can reconstruct depth image from color image accurately and efficiently. Despite many existing work on shape recovery from single image, most of them rely on 3D model fitting which is time-consuming and can be prone to lack accuracy when landmarks are not precisely located. Thanks to the extremely rapid development of generative models, especially the Generative Adversarial Network (GAN) [10] and its conditional variation (cGAN) [22] which are introduced quite recently, we implement an end-to-end depth face recovery with cGAN to enforce the realistic image generation. Furthermore, the recovered depth information enables a straightforward comparison in 2.5D space.

A flowchart of the proposed method is illustrated in Fig. 1, and we list our contributions as follows:

- A novel depth face recovery method based on cGAN and Auto-encoder with skip

connections which greatly improves the quality of reconstructed depth images.

- We first train two discriminative CNNs individually for a two-fold purpose: to extract features of color image and depth image, and to provide pre-trained models for the cross-modal 2D/2.5D CNN model.

- A novel heterogeneous face recognition pipeline which fuses multi-modal matching scores to achieve state-of-the-art performance.

2 Related Work

2.1 3D Face Reconstruction

3D face reconstruction from single/multiple images or stereo video has been a challenging task due to its nonlinearity and ill-posedness. A number of prevailing approaches addressed this problem based on shape-subspace projections, where a set of 3D prototypes are fitted by adjusting corresponding parameters to a given 2D image and most of them were derived from 3DMM [4] and Active Appearance Models [21]. Alternative models were afterwards proposed as well which follow the similar processing pipeline by fitting 3D models to 2D images through various face collections or prior knowledge. For example, Gu and Kanade [11] fit surface 3D points and related textures together with the pose and deformation estimation. Kemelmacher-Shlizerman et al. [17] considered the input image as a guide with a single reference model to achieve 3D reconstruction. In recent work of Liu et al. [20], two sets of cascaded regressors are implemented and correlated via a 3D-2D mapping iteratively to solve face alignment and 3D face reconstruction simultaneously. Likewise, using generic model remains a decent solution as well for 3D face reconstruction from stereo videos, as presented in [6, 9, 23]. Despite of strikingly accurate reconstruction result reported in the above researches, the drawback of relying on single or a large number of well-aligned 3D training data is observed and even enlarged here, because as far as we know 3D prototypes are necessary for almost all reconstruction approaches.

2.2 2D-3D Heterogeneous Face Recognition

As a pioneer and cornerstone for numerous subsequent 3D Morphable Model (3DMM) based methods, Blanz and Vetter [4] built this statistical model by merging a branch of 3D face models and then densely fit it to a given facial image for further matching. Toderici et al. [33] located some predefined key landmarks on the facial images in different poses, and then roughly align them to a frontal 3D model to achieve recognition target; Riccio and Dugelay [26] also established a dense correspondence between the 2D probe and the 3D gallery using geometric invariants across face region. Following this framework, a pose-invariant asymmetric 2D-3D FR approach [39] was proposed which conducts a 2D-2D matching by synthesizing 2D image from corresponding 3D models towards the same pose as a given probe sample. This approach was further extended and compared with work of Zhao et al. [41] as a benchmarking asymmetric 2D-3D FR system, a complete version of their work was recently released in [16]. Though the above models achieved satisfactory performance, unfortunately they all suffer from high computational cost and long convergence process owing to considerable complexity of pose synthesis, and their common assumption that accurate landmark localization in facial images was fulfilled turns out to be another tough topic. More

recently, learning based approaches have significantly increased on 2D/3D FR. Huang et al. [13] projected the proposed illuminant-robust feature OGM onto the CCA space to maximize the correlation between 2D/3D features; instead, Wang et al. [35] combined Restricted Boltzmann Machines (RBMs) and CCA/kCCA to achieve this goal. The work of Jin et al. [15], called MSDA based on Extreme Learning Machine (ELM) as aforementioned, aims at finding a common discriminative feature space revealing the underlying relationship between different views. These approaches take well advantage of learning model, but would encounter weakness when dealing with non-linear manifold representations.

3 Depth Face Reconstruction

The target of taking a random color face image to recover its counterpart in depth space is realized in this section. We first formulate our problem by adapting it to the background of cGAN, then the detailed architecture design is described and discussed.

3.1 Problem Formulation

First proposed in [10], GAN has achieved impressive results in a wide variety of generative tasks. The core idea of GAN is to train two neural networks, which respectively represent the generator G and the discriminator D, to proceed a game-theoretic tussle between one another. Given the samples x from the real data distribution $p_{data}(x)$ and random noise z sampled from a noise distribution $p_z(z)$, the discriminator aims to distinguish between real samples x and fake samples which are mapped from z by the generator, while the generator is tasked with maximally confusing the discriminator. The objective can thus be written as:

$$\mathcal{L}_{GAN}(G,D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

where $\mathbb{E}$ denotes the empirical estimate of expected value of the probability. To optimize this loss function, we aim to minimize its value for G and maximize it for D in an adversarial way, i.e. $\min_G \max_D \mathcal{L}_{GAN}(G,D)$.

The advantage of GAN is that realistic images can be generated from noise vectors with random distribution, which is crucially important for unsupervised learning. However, note that in our face recovery scenario, training data contains image pairs $\{x,y\}$ where x and y refer to the depth and color faces respectively with a one-to-one correspondence between them. The fact that y can be involved in the model as a prior for generative task leads us to the conditional variant of GAN, namely cGAN [22]. Specifically, we condition the observations y on both the discriminator and the generator, the objective of cGAN extends (1) to:

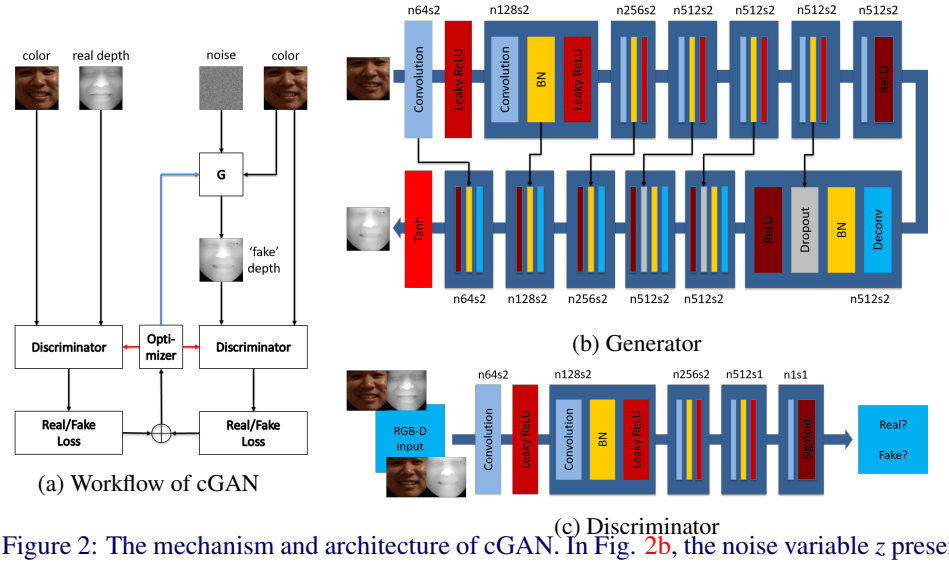
$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y\sim p_{data}(x,y)}[\log D(x,y)] + \mathbb{E}_{z\sim p_z(z),y\sim p_{data}(y)}[\log(1-D(G(z|y),y))] \tag{2}$$

Moreover, to ensure the pixel-wise similarity between image generation outputs $G(z|y)$ and ground truth x, we subsequently impose a reconstruction constraint on the generator in the form of L1 distance between them:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y\sim p_{data}(x,y),z\sim p_z(z)}[\|x-G(z|y)\|_1] \tag{3}$$

The comprehensive objective is formulated with a minmax value function on the above two losses where the scalar η is used for balancing them:

$$\min_G \max_D [\mathcal{L}_{cGAN}(G,D) + \eta \mathcal{L}_{L1}(G)] \tag{4}$$



(a) Workflow of cGAN

(b) Generator

(c) Discriminator

Figure 2: The mechanism and architecture of cGAN. In Fig. 2b, the noise variable z presents itself under the the form of dropout layers, while the black arrows portray the skip connections. All convolution and deconvolution layers are with filter size 4×4 and 1-padding, n and s represent the number of output channels and stride value, respectively. (Best view in color)

Note that the cGAN itself can hardly generate specified images and using only $\mathcal{L}_{L1}(G)$ causes blurring, this joint loss successfully leverages the complementary strengths of them.

3.2 CGAN Architecture

We adapt our cGAN architecture from that in [14] which achieved particularly impressive results in image-to-image translation task. A detailed description of this model is illustrated in Fig. 2 and some key features are discussed below.

Generator: As a standard generative model, the architectures of auto-encoder (AE) [12] and its variants [18, 27, 34] are widely adopted as G for past cGANs. However, the drawback of conventional AEs is obvious: due to their dimensionality reduction capacity, a large portion of low-level information, such as precise localization, is compressed when an image passes through layers in the encoder. To cope with this lossy compression problem, we follow the idea of U-Net [28] by adding skip connections which forwards directly the features from encoder layers to decoder layers that are on the same 'level', as shown in Fig. 2b.

Discriminator: Consistent with Isola et al. [14], we adopt *Patch*GAN for the discriminator. Within this pattern, no fully connected layers are implemented and D outputs a 2D image where each pixel represents the prediction result with respect to the corresponding patch on original image. All pixels are then averaged to decide whether the input image is 'real' or 'fake'. Compared with pixel-level prediction, *Patch*GAN efficiently concentrates on local patterns while the global low-frequency correctness is enforced by L1 loss in (3).

Optimization: The optimization for cGAN is performed by following the standard method [10]: the mini-batch SGD and the Adam solver are applied to optimize G and D alternately (as depicted by arrows with different colors in Fig. 2a).

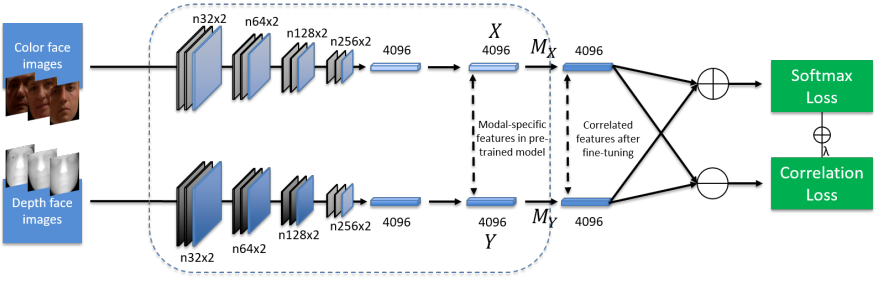


Figure 3: Training procedure of the cross-modal CNN model. Models in the dashed box are pre-trained using 2D and 2.5D face images individually.

4 Heterogeneous Face Recognition

The reconstruction of depth faces from color images enables us to maximally leverage shape information in both gallery and probe, which means we can individually learn a CNN model to extract discriminative features for depth images and transform the initial cross-modal problem into a multi-modal one. However, the heterogeneous matching remains another challenge in our work, below we demonstrate how this problem is formulated and tackled.

Unimodal learning. The last few years witnessed a surge of interest and success in FR with deep learning [24, 30, 31]. Following the basic idea of stacking convolution-convolution-pooling (C-C-P) layers in [19], we train from scratch two CNNs for color and grayscale images on CASIA-WebFace [37] and further fine-tune the grayscale based model with our own depth images. These two models serve two purposes: to extract 2D and 2.5D features individually, and to offer pre-trained models for the ensuing cross-modal learning.

Cross-modal learning. Once a pair of unimodal models for both views are trained, the modal-specific representations, $\{X, Y\}$, can be obtained after the last fully connected layers. Note that each input for the two-stream cross-modal CNN is a 2D+2.5D image pair with identity correspondence, it is reasonable to have an intuition that X and Y share common patterns which help to classify them as the same class. This connection essentially reflects the nature of cross-modal recognition, and was investigated in [13, 35, 36].

In order to explore this shared and discriminative feature, a joint supervision is required to enforce both correlation and distinctiveness simultaneously. For this purpose, we apply two linear mappings following X and Y, denoted by M_X and M_Y. First, to ensure the correlation between new features, they are enforced to be as close as possible, which is constrained by minimizing their distance in feature space:

$$\mathcal{L}_{corr} = \sum_{i=1}^{n} \|M_X X_i - M_Y Y_i\|_F^2 \tag{5}$$

where n denotes the size of mini-batch and $\|\cdot\|_F$ represents the Frobenius norm.

If we only use the above loss supervision signal, the model will simply learn zero mappings for M_X and M_Y because the correlation loss will stably be 0 in this case. To avoid this tricky situation, we average the two outputs to obtain a new feature on which the classifica-

Databases	Training Set			Test set
	BU3D [38]	Bosphorus [29]	CASIA-3D [1]	FRGC Ver2.0 [25]
# Persons	100	105	123	466
# Images	2500	2896	1845	4003
Conditions	E	E	EI	EI

Table 1: Database overview. E and I are short for expressions and illuminations, respectively.

tion loss is computed. The ultimate objective function is formulated as follows:

$$\begin{aligned}\mathcal{L}_{hfr} &= \mathcal{L}_{softmax} + \lambda \mathcal{L}_{Corr} \\ &= -\sum_{i=1}^{n} \log \frac{e^{W_{c_i}^T (M_X X_i + M_Y Y_i)/2 + b_{c_i}}}{\sum_{j=1}^{m} e^{W_j^T (M_X X_i + M_Y Y_i)/2 + b_j}} + \lambda \sum_{i=1}^{n} \|M_X X_i - M_Y Y_i\|_F^2\end{aligned}$$

where c_i represents the ground truth class label of ith image pair, the scalar λ denotes the weight for correlation loss.

Fusion. To highlight the effectiveness of the proposed method, we adopt the cosine similarity of 4096-d hidden layer features as matching scores. As for the score fusion stage, all scores are normalized to [0,1] and fused by a simple sum rule.

5 Experimental Results

To intuitively demonstrate the effectiveness of the proposed method, we conduct extensive experiments for 2D/2.5D HFR on the benchmark 2D/2.5D face database. Besides the reconstructed 2.5D depth image, our method also outperforms state-of-the-art performance using only 2.5D images instead of holistic 3D face models.

5.1 Dataset Collection

Collecting 2D/2.5D image pairs presents itself as a primary challenge when considering deep CNN as a learning pipeline. Unlike the tremendous boost in dataset scale of 2D face images, massive 3D face data acquisition still remains a bottleneck for the development and practical application of 3D based FR techniques, from which our work is partly motivated.

Databases: As listed in Table 1, three large scale and publicly available 3D face databases are gathered as training set and the performance is evaluated on another dataset, which implies that there was no overlap between training and test set and the generalization capacity of the proposed method is evaluated as well. Note that the attribute values only concern the data used in our experiments, for example, scans with large pose variations in CASIA-3D are not included here.

Preprocessing: To generate 2.5D range image from original 3D shape, we either proceed a direct projection if the point cloud is pre-arranged in grids (Bosphorus/FRGC) or adopt a simple Z-buffer algorithm (BU3D/CASIA3D). Furthermore, to ensure that all faces are of the similar scale, we resize and crop the original image pairs to 128×128 while fixing their interocular distance to a certain value. Especially, to deal with the missing holes and unwanted body parts (shoulder for example) in raw data of FRGC, we first locate the face based on 68 automatically detected landmarks [2], and then apply a linear interpolation to approximate the default value of each hole pixel by averaging its non-zero neighboring points.

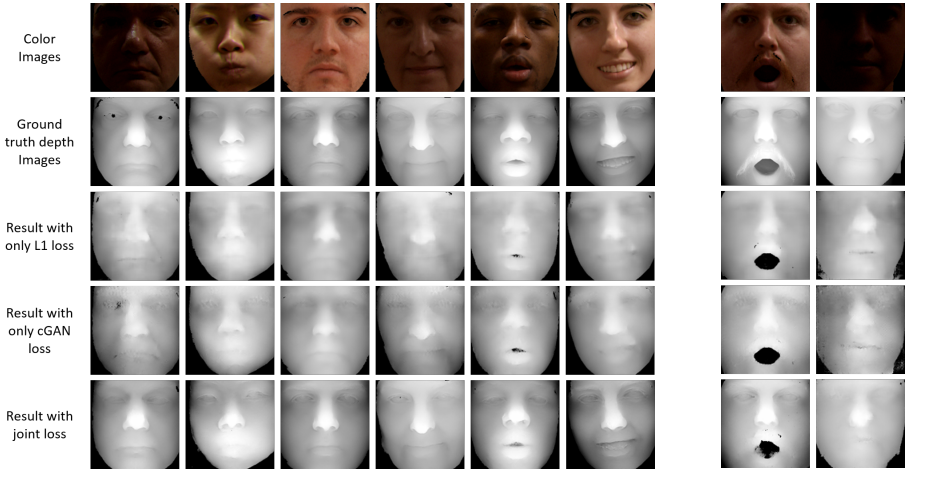


Figure 4: Qualitative reconstruction results of FRGC samples with varying illuminations and expressions. Fig. 4a: correctly recovered samples. Fig . 4b: wrongly recovered samples.

5.2 Implementation details

All images are normalized before being fed to the network by subtracting from each channel its mean value over all training data. With regards to the choice of hyperparameters, we adopt the following setting: in cGAN, the learning rate μ_{cGAN} is set to 0.0001 and the weight for L1 norm η is 500; in cross-modal CNN model, the learning rate for training from scratch μ begins with 1 and is divided by 5 every 10 epochs while the learning rate during fine-tuning μ_{ft} is 0.001; for both models, the momentum m is initially set as 0.5 until it is increased to 0.9 at the 10th epoch; the weight for correlation loss λ is set to 0.6.

5.3 Reconstruction Results

The reconstruction results obtained for color images in FRGC are illustrated in Fig. 4. Samples from different subjects across expression and illumination variations are shown from left to right. They thereby give hints on the generalization ability of the proposed method. For each sample we first portray the original color image with its ground truth depth image, followed by the reconstructed results whereby we demonstrate the effectiveness and necessity of each constraint in the joint objective. In addition, some samples with low reconstruction quality are depicted in Fig. 4b as well.

Being consistently similar with the ground truth, the reconstruction results with joint loss in Fig. 4a intuitively demonstrate the strength of cGAN. The recovered depth faces hold their accuracy and realistic property irrespective of lighting and expression variations in the original RGB images. Furthermore, when we take an observation of the two reconstruction results in 3rd and 4th rows, the comparison implies that: 1) using only L1 loss will lead to blurry results because the model tends to average all plausible values, especially for regions containing high-level information like edges; 2) using only cGAN loss can achieve slightly sharper results, but suffers from noise. These results provide an evidence that the implementation of joint loss is beneficial and important for obtaining a 'true' and accurate output. Meanwhile, our model encounters some problems while dealing with extreme cases, such as

Protocol	Methods	Rank-1 Recognition Accuracy			
		2D	2.5D	2D/2.5D	Fusion
Jin et al. [15]	MSDA+ELM [15]	-	-	0.9680	-
	Ours	-	0.9573	0.9603	**0.9698**
Wang et al. [35]	GRBM+rKCCA [35]	-	-	0.9600	-
	Ours	-	0.9529	0.9714	**0.9745**
Huang et al. [13]	OGM [13]	0.9390	-	0.9404	0.9537
	Ours	0.9755	0.9609	0.9688	**0.9792**

Table 2: Comparison of recognition accuracy on FRGC under different protocols.

λ	0	0.2	0.4	0.6	0.8	1	1.2
Accuracy	0.9245	0.9481	0.9600	0.9688	0.9577	0.8851	0.7333

Table 3: 2D/2.5D HFR accuracy with varying λ under protocol of [13].

thick beard, wide opened mouth and extremely dark shadows as displayed in Fig. 4b. The errors are principally due to few training samples with these cases.

5.4 2D-3D Asymmetric FR

We conduct the quantitative experiments on FRGC which has held the field as one of the most commonly used benchmark dataset over the last decade. In contrast with unimodal FR experiments, very few attempts have been made on 2D/3D asymmetric FR. For convenience of comparison, three recent and representative protocols reported respectively in [15], [13] and [35] are followed. These protocols mainly differ in gallery and probe setting, including splitting and modality setup. For example, the gallery set in [35] solely contains depth images, when compared with their work, our experiment will subsequently exclude 2D based matching to respect this protocol.

The comparison results are shown in Table 2, through which we could gain the observation that the proposed cross-modal CNN outperforms state-of-the-art performance while fusing 2.5D matching into HFR with reconstructed depth image further helps improve the performance effectively. Moreover, the proposed method is advantageous in its 3D-free reconstruction capacity and efficiency. To the best of our knowledge, this is the first time to investigate a 2.5D face recovery approach which is free of any 3D prototype models. Despite nearly 20 hours for the whole training and fine-tuning procedure, it takes only 1.6 *ms* to complete an online forward pass per image on a single NVIDIA GeForce GTX TITAN X GPU and is therefore capable of satisfying the real-time processing requirement.

Effect of hyperparameter λ. An extended analysis is made to explore the role of softmax loss and correlation loss. We take the protocol in [13] as a standard and vary the weight for correlation loss λ each time. As shown in Table 3, the performance will remain largely stable across a range of λ_c between 0.4 and 0.8. When we set $\lambda = 0$ instead of 0.6, which means correlation loss is not involved while training, the network can still learn valuable features with a recognition rate decrease of 4.43%. However, along with the increase of λ, the performance drops drastically, which implies that a too strong constraint on correlation loss could backfire by causing a negative impact on softmax loss.

6 Conclusion

In this paper, we have presented a novel framework for 2D/2.5D heterogeneous face recognition together with depth face reconstruction. This approach combines the generative capacity of conditional GAN and the discriminative feature extraction of deep CNN for cross-modality learning. The extensive experiments have convincingly evidenced that the proposed method successfully reconstructs realistic 2.5D from single 2D while being adaptive and sufficient for HFR. This architecture could hopefully be generalized to other heterogeneous FR tasks, such as visible light vs. near-infrared and 2.5D vs. forensic sketch, which provides an interesting and promising prospect.

7 Acknowledgement

This work was supported in part by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the Jemime project (N° contract ANR-13-CORD-0004-02), the Biofence project (N° ANR-13-INSE-0004-02) and the PUF 4D Vision project funded by the Partner University Foundation.

References

[1] CASIA-3D FaceV1 database. http://biometrics.idealtest.org/, 2004.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1866, 2014.

[3] Aisha Azeem, Muhammad Sharif, Mudassar Raza, and Marryam Murtaza. A survey: Face recognition techniques under partial occlusion. *Int. Arab J. Inf. Technol.*, 11(1): 1–10, 2014.

[4] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003.

[5] Kevin W Bowyer, Kyong Chang, and Patrick Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1):1–15, 2006.

[6] A Roy Chowdhury, Rama Chellappa, Sandeep Krishnamurthy, and Tai Vo. 3d face reconstruction from video using a generic model. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 449–452. IEEE, 2002.

[7] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[8] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):37, 2016.

[9] Douglas Fidaleo and Gérard Medioni. Model-assisted 3d face reconstruction from video. In *International Workshop on Analysis and Modeling of Faces and Gestures*, pages 124–138. Springer, 2007.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] Lie Gu and Takeo Kanade. 3d alignment of face in a single image. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 1305–1312. IEEE, 2006.

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[13] Di Huang, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Oriented gradient maps based automatic asymmetric 3d-2d face recognition. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 125–131. IEEE, 2012.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[15] Yi Jin, Jiuwen Cao, Qiuqi Ruan, and Xueqiao Wang. Cross-modality 2d-3d face recognition via multiview smooth discriminant analysis based on elm. *Journal of Electrical and Computer Engineering*, 2014:21, 2014.

[16] Ioannis A Kakadiaris, George Toderici, Georgios Evangelopoulos, Georgios Passalis, Dat Chu, Xi Zhao, Shishir K Shah, and Theoharis Theoharis. 3d-2d face recognition with pose and illumination normalization. *Computer Vision and Image Understanding*, 2016.

[17] Ira Kemelmacher-Shlizerman and Ronen Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):394–405, 2011.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[19] Yuancheng Lee, Jiancong Chen, Ching-Wei Tseng, and Shang-Hong Lai. Accurate and robust face recognition from rgb-d images with a deep learning approach. In *BMVC*, 2016.

[20] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *European Conference on Computer Vision*, pages 545–560. Springer, 2016.

[21] Iain Matthews, Jing Xiao, and Simon Baker. 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *International journal of computer vision*, 75(1):93–113, 2007.

[22] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[23] Unsang Park and Anil K Jain. 3d model-based face recognition in video. In *International Conference on Biometrics*, pages 1085–1094. Springer, 2007.

[24] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[25] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954. IEEE, 2005.

[26] Daniel Riccio and Jean-Luc Dugelay. Geometric invariants for 2d/3d face recognition. *Pattern Recognition Letters*, 28(14):1907–1914, 2007.

[27] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 833–840, 2011.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[29] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[30] Yi Sun, Ding Liang, Xiaogang Wang, and Xiaoou Tang. Deepid3: Face recognition with very deep neural networks. *arXiv preprint arXiv:1502.00873*, 2015.

[31] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[32] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern recognition*, 39(9):1725–1745, 2006.

[33] George Toderici, Georgios Passalis, Stefanos Zafeiriou, Georgios Tzimiropoulos, Maria Petrou, Theoharis Theoharis, and Ioannis A Kakadiaris. Bidirectional relighting for 3d-aided 2d face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2721–2728. IEEE, 2010.

[34] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11 (Dec):3371–3408, 2010.

[35] Xiaolong Wang, Vincent Ly, Rui Guo, and Chandra Kambhamettu. 2d-3d face recognition via restricted boltzmann machines. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, pages 574–580. IEEE, 2014.

[36] Ziyan Wang, Ruogu Lin, Jiwen Lu, Jianjiang Feng, et al. Correlated and individual multi-modal deep learning for rgb-d object recognition. *arXiv preprint arXiv:1604.01655*, 2016.

[37] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[38] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[39] Wuming Zhang, Di Huang, Yunhong Wang, and Liming Chen. 3d aided face recognition across pose variations. In *Chinese Conference on Biometric Recognition*, pages 58–66. Springer, 2012.

[40] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[41] Xi Zhao, Wuming Zhang, Georgios Evangelopoulos, Di Huang, Shishir K Shah, Yunhong Wang, Ioannis A Kakadiaris, and Liming Chen. Benchmarking asymmetric 3d-2d face recognition systems. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013.