GridFace: Face Rectification via Learning Local Homography Transformations

Erjin Zhou, Zhimin Cao, and Jian Sun

Face++, Megvii Inc.
{zej,czm,sunjian}@megvii.com

Abstract. In this paper, we propose a method, called GridFace, to reduce facial geometric variations and improve the recognition performance. Our method rectifies the face by local homography transformations, which are estimated by a face rectification network. To encourage the image generation with canonical views, we apply a regularization based on the natural face distribution. We learn the rectification network and recognition network in an end-to-end manner. Extensive experiments show our method greatly reduces geometric variations, and gains significant improvements in unconstrained face recognition scenarios.

Keywords: Face Recognition · Face Rectification · Homography Transformation

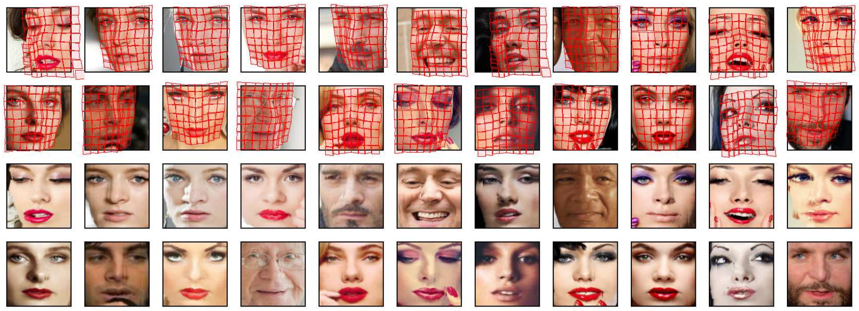


Fig. 1: Face Rectification Examples. The top two rows are faces with large geometric variations, and local homographies estimated by the rectification network. The bottom two rows show the rectified faces by local homographies, which greatly reduce the geometric variations and calibrate the faces into canonical view.

1 Introduction

Despite of the recent academic/commercial progresses made in deep learning [34, 31, 30, 47, 41, 28, 18, 37, 20, 36, 14, 42, 43, 39], it is still hard to claim that face recognition has been solved in unconstrained settings. One of the remaining challenges for the in-the-wild recognition is facial geometric variations. Such variations in pose and misalignment (introduced by face detection bounding

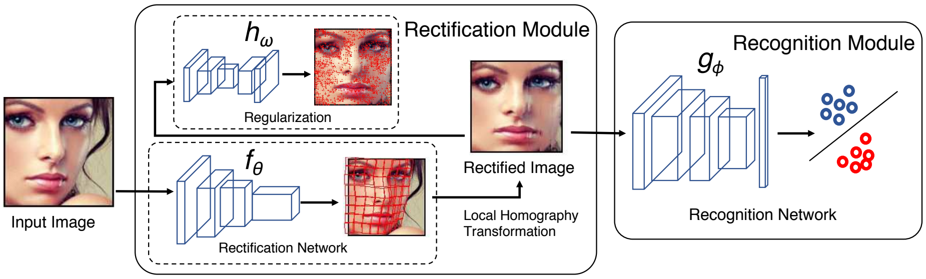


Fig. 2: System Overview. The system contains two modules: the rectification module and the recognition module. The rectification module extracts deep feature by the rectification network and warps the image with a group of local homography transformations (Sec. 3.2). The rectified output is regularized by an implicit canonical view face prior, which is optimized by a Denoising Autoencoder (Sec. 3.3). The red arrows in the face in the regularization box indicate the approximated gradients estimated by DAE. With the rectified faces as input, the recognition network learns discriminative face representation (Sec. 3.4) via metric learning. The whole system is end-to-end optimized with stochastic gradient descent.

box localization) substantially degrade the face representation and recognition performance.

The common adopted way to deal with this issue is using a 2D transformation to calibrate the facial landmarks to pre-defined templates (i.e., 2D mean face landmarks or a 3D mean face model). However, such kind of pre-processing is not optimized towards the recognition system and relies heavily on the parameters tuned by hand and accurate facial landmarks. To address this problem, recent works use the Spatial Transformer Network (STN) [15] to perform an end-to-end optimization with consideration of both face alignment and detection/recognition [5, 44]. However, the transformation learned in these works uses a holistic parametric model that can only capture coarse geometric information, such as facial orientation, and may introduce notable distortion in the rectified results.

In this paper, we propose a novel method called *GridFace* to reduce the facial geometric variations and boost the recognition performance. As shown in Fig. 2, our system contains two main modules: the rectification module and the recognition module.

In the rectification module, we apply a face rectification network to estimate a group of local homography transformations for rectifying the input facial image (Sec. 3.2). We approximate the underlying 3D canonical face shape by a group of deformable plane cells. When a face with geometric variations fed in, local homography transformations are estimated to model the warping of each cell respectively. In order to encourage the generation with canonical views, we introduce a regularization based on the canonical view face distribution (Sec. 3.3). This natural face distribution is not explicitly modeled. Instead, we use a Denoising Autoencoder (DAE) to estimate the gradients of logarithm of probability density, which is inspired by the previous work [27, 1]. The recognition module (Sec. 3.4) takes the rectified image as input and learns discriminative representation via metric learning.

In Sec. 4, we first evaluate our method with qualitative and quantitative results to demonstrate the effectiveness of face rectification for recognition in-the-wild. Then we present extensive ablation studies to show the importance of each of the above components. We finally evaluate our method on four challenging public benchmarks LFW, YTF, IJB-A, and Multi-PIE. We obtain large improvement in all benchmarks, and achieve superior or comparable results compared with recent face frontalization and recognition works.

Our contributions are summarized as following:

1. We propose a novel method to improve face recognition performance by reducing facial geometric variations with local homography transformations.
2. We introduce a canonical face prior and a Denoising Autoencoder based approximation method to regularize the face rectification process for better rectification quality.
3. Extensive experiments on constrained and unconstrained environments are conducted to demonstrate the excellent performance of our method.

2 Related Works

Deep Face Recognition. Early works [31, 34] learn face representation by multi-class classification networks. Features learned from thousands of individuals' faces demonstrate good generalization ability. Sun et al. [30] improve the performance by jointly learning identification and verification losses. Schroff et al. [28] formulate the representation learning task in a metric learning framework, and introduce the triplet loss and hard negative sample mining skill to boost the performance further. Recent works [37, 18] propose the center loss and sphere loss to further reduce intra-class variations in the feature space. Du and Liang [8] propose age-invariant feature. Bhattarai et al. [3] introduce multitask learning for large scale face retrieval. Zhang et al. [43] develop a range loss to effectively utilize the long tail training data. Pose invariant representation is the key step for real world robust recognition system, and has been the focus of many works. For example, Masi et al. [20] propose the face representation by fusing multiple pose-aware CNN models. Peng et al. [25] untangle the identity and pose in representation by reconstruction in the feature space. Lu et al. [19] propose the joint optimization framework for face and pose tasks.

Face Frontalization and Canonicalization. Prior works in face frontalization and canonicalization optimize an image warping to fit a 3D face model [45, 12] based on localized 2D facial landmarks. Recently, several attempts have been made to improve the generated face quality with neural networks. Early works [46],[47] calibrate faces of various poses into canonical view, and disentangle the pose factor from identity with convolution neural networks. Yim et al. [41] improve the identity preserving ability by introducing an auxiliary task to reconstruct the input data. Cole et al. [6] decompose the generation module into geometry and texture parts, training with the differentiable warping.

Recent works further improve the generation quality with Generative Adversarial Network (GAN) [9]. Tran et al. [36] propose DR-GAN to simultaneously learn the frontal face generation and discriminative representation disentangled from pose variations. Yin et al. [42] introduce a 3DMM reconstruction module in the proposed FF-GAN framework to provide better shape and appearance prior. Huang et al. [14] incorporate both global structure and local details in their generator with landmark located patch networks. In our method, we do not require frontal and profile training pairs that are needed in the previous work, and our rectification process is recognition oriented, which induces better recognition performance.

Spatial Transformer Network. The Spatial Transformer Network (STN) [15] performs spatial transforms in the image or feature maps with a differential module, which can be integrated into the deep learning pipeline and optimized end-to-end . The most relevant application of STN to our work is image alignment. Kanazawa et al. [16] match the fine-grained objects by establishing correspondences between two input images with non-rigid deformations. Chen et al. [5] use STN to warp face proposals to canonical view with detected facial landmarks. Zhong et al. [44] use STN for face alignment before recognition. Lin et al. [17] provide a theoretical connection between STN and the Lucas-Kanade algorithm, and introduce the the inverse composition STN to reduce input variations.

The recent work Wu et al. [39] propose a recursive spatial transformer (ReST) for the alignment-free face recognition. They also integrate the recognition network in an end-to-end optimization manner. There are two major differences between our approach and ReST. First, instead of manually dividing the facial region into several regions to allow non-rigid transformation modeling, we use a group of deformable plane cells to deal with complex warping effects. Second, we introduce a regularization prior of canonical view face to achieve better rectification effects.

3 Approach

Notation. Let I^X, I^Y denote the original image and rectified image. We define the coordinate in the original image I^X as the *original coordinate*, and the one in the rectified image I^Y as the *rectified coordinate*. Let $p = [p_x, p_y]^T$ and $q = [q_x, q_y]^T$ denote the points in the original coordinate and rectified coordinate. We use $\hat{p}$ and $\hat{q}$ to denote the homogeneous coordinates as $\hat{p} = [p_x, p_y, 1]^T, \hat{q} = [q_x, q_y, 1]^T$. Without loss of generality, we assume the coordinates of pixels are normalized to $[0, 1) \times [0, 1)$.

3.1 Overview

The system contains two parts: the rectification module and the recognition module (Fig. 2). In the rectification process, the rectification network f_θ with

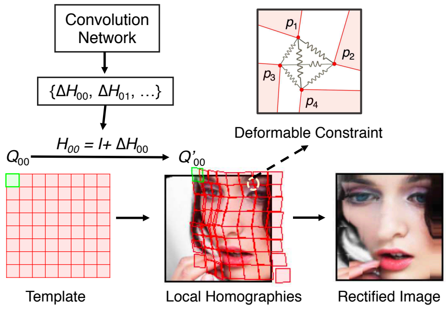


Fig. 3: Local Homography Transformation. The rectification process approximates the 3D face as plane cells and canonicalizes it with local homographies. The rectified image is partitioned into n^2 cells, and the corresponding homographies are estimated by the rectification network. We put springs at the corners of the cells as soft constraints to avoid large discontinuities in the boundaries.

parameter θ maps the original face image I^X into the rectified one I^Y by non-rigid image warping. Then, the recognition network g_ϕ is trained with the metric learning based on the rectified image I^Y. We further introduce a regularization to encourage the rectified face in canonical views, which is modeled as a prior under the distribution of natural faces with canonical views.

3.2 Face Rectification Network

In this section, we present the rectification process. Different from recent face frontalization techniques [36, 42, 14] generating faces from abstract feature, we define the rectification process as warping pixels from the original image to the canonical one, as illustrated in Fig. 3.

Formally, we define a template Q by partitioning the rectified image into n^2 non-overlapped cells

$$\begin{aligned} Q &= \{Q_{i,j}\}, 0 \le i, j < n, \\ Q_{i,j} &= \left[\frac{i}{n}, \frac{i+1}{n}\right) \times \left[\frac{j}{n}, \frac{j+1}{n}\right). \end{aligned} \tag{1}$$

For each cell $Q_{i,j}$, we compute the corresponding deformed cell $Q'_{i,j}$ in the original image by estimating a local homography $H_{i,j}$.

Specifically, we formulate the homography matrix as

$$H_{i,j} = \begin{pmatrix} 1+h_1 & h_2 & h_3 \\ h_4 & 1+h_5 & h_6 \\ h_7 & h_8 & 1 \end{pmatrix} = I + \mathit{\Delta} H_{i,j} \tag{2}$$

The rectification network takes the original image I^X as input and predicts n^2 residual matrices $\mathit{\Delta} H_{i,j}$. Then the rectified image I^Y at cell $Q_{i,j}$ is obtained with homographies $H_{i,j} = I + \mathit{\Delta} H_{i,j}$ as

$$I^Y_q = I^X_p, q \in Q_{i,j}, p \in Q'_{i,j}, \quad \text{s.t.} \quad \lambda \hat{p} = H_{i,j}\hat{q}, \lambda \ne 0, \tag{3}$$

where $\hat{p}, \hat{q}$ are the homogeneous coordinates of p, q.

Let C denote the collection of corners of each cell $Q_{i,j}$ as $C = \{(\frac{i}{n}, \frac{j}{n}); 0 \leq i, j \leq n\}$. Since all the local homographies are estimated separately, a cell corner $c_i \in C$ in the rectified image is mapped to multiple points in the original image (see Fig. 3). In order to avoid large discontinuities between the boundaries of neighboring cells in I_X, we further introduce a soft constraint, called deformation constraint $\mathcal{L}_{\text{de}}$. Specifically, let M_i denotes the collection of c_i's mapping coordinates in the original image. Then a soft constraint $\mathcal{E}_{c_i}$ is added to enforce the conformity between every pair of points in M_i as $\mathcal{E}_{c_i} = \sum_{u,v \in M_i} ||u - v||_2$. We incorporate this soft constraint into the learning objective, and cast it as the the deformation loss of the rectification network:

$$\mathcal{L}_{de} = \frac{1}{|C|} \sum_{c_i \in C} \mathcal{E}_{c_i}. \tag{4}$$

3.3 Regularization by Denoising Autoencoder

The regularization encourages that the rectification process generates face in canonical views. We define it as an image prior that is directly based on the natural canonical view face distribution P_Y as

$$\mathcal{L}_{reg} = -\log P_Y(I^Y). \tag{5}$$

In general, this optimization is non-trivial. We do not explicitly model the distribution, but consider the gradient of $\log P_Y$ and maximize it with stochastic gradient descent

$$\frac{\partial}{\partial \theta} \log P_Y(I^Y) = \frac{\partial}{\partial I^Y} \log P_Y(I^Y) \frac{\partial I^Y}{\partial \theta}. \tag{6}$$

Using results from [27],[1], which are also used in image generation [23] and restoration [29],[4],[22], we approximate the gradient of the prior as

$$\frac{\partial}{\partial I^Y} \log P_Y(I^Y) \approx \frac{h_{\omega^*}(I^Y) - I^Y}{\sigma^2}. \tag{7}$$

Here $h_{\omega^*} = \text{argmin}_{h_\omega} E_y ||h_\omega(y + \sigma\epsilon) - y||_2^2$, with $\epsilon \sim N(0, I)$ and $y \sim P_Y$, is the optimal denoising autoencoder trained on the true data distribution P_Y (canonical view faces in our work) with the infinitesimal noise level σ. Using these results, we optimize the Eqn. 5 by first training a Denoising Autoencoder h_ω on the canonical view face dataset, and then estimating the approximated gradient in backpropagation via Eqn. 7.

3.4 Face Recognition Network

Given the rectified face $f_\theta(I^X) = I^Y$, we extract the face representation $g_\phi(I^Y)$ with deep convolutional recognition network g_ϕ. Following the previous works [28], we train the recognition network with triplet loss. Let $D = \{I_o^X, I_+^X, I_-^X\}$ denote

Network	**Rectification Network**	**Denoising Autoencoder**
Input	128 × 128 × 3	
Stage-1	Conv[8, 3, 2, 1]	
	MaxPool[2,2,1]	Conv[8, 3, 2, 1]
	Conv[32, 3, 2, 1]	Conv[12, 3, 2, 1]
Stage-2	InceptionModule[16]	Conv[16, 3, 2, 1]
	MaxPool[2, 2]	Conv[24, 3, 2, 1]
Stage-3	InceptionModule[32]*2	FullyConnected[1536]
	MaxPool[2, 2]	DeConv[24, 3, 2, 1]
Stage-4	InceptionModule[64]*2	DeConv[16, 3, 2, 1]
	MaxPool[2, 2]	DeConv[12, 3, 2, 1]
Stage-5	FullyConnected[128]	DeConv[8, 3, 2, 1]
	FullyConnected[N]	Conv[3, 3, 1, 1]

Table 1: Network Details. Conv$[ch, w, s, p]$ denotes a convolution layer with kernel size $ch \times w \times w$, stride s and padding p. The deconvolution layer DeConv$[ch, w, s, p]$ is implemented as the gradient of convolution with respect to data, and the meaning of parameters is still in a convolution sense. MaxPool$[w, s]$ is a max-pooling layer with $w \times w$ window and stride s. FullyConnected$[n]$ is a fully-connected layer with n output neurons, and N denotes the number of corresponding transformation parameters. InceptionModule$[ch]$ denotes a modified Inception module with the same number of feature maps ch in each branch.

the three images forming a face triplet where I_o^X and I_+^X are from the same person, while I_-^X is from a different person. the recognition loss is

$$\mathcal{L}_{recog} = \max\Big(0, d\big(I_o^X, I_+^X\big) - d\big(I_o^X, I_-^X\big) + \alpha\Big). \tag{8}$$

where $d(x, y) = ||g_\phi(f_\theta(x)) - g_\phi(f_\theta(y))||_2$ is the Euclidean distance between the feature representations x and y. The hyper-parameters α control the margin between intra-person distance and inter-person distance in the triplet loss.

In summary, we jointly optimize the rectification network and recognition network by minimizing an objective, consisting of a deformable term, a recognition term, and a regularization term

$$\underset{\theta,\phi}{\operatorname{argmin}} \quad \mathcal{L}_{\text{recog}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{de}}\mathcal{L}_{\text{de}}. \tag{9}$$

4 Experiments

4.1 Experimental Details

Dataset. Our models are learned from a large collection of photos in social networks, referred to as the Social Network Face Dataset (SNFace). The SNFace dataset contains about 10M images and 200K individuals. We randomly choose 2K individuals as the validation set, 2K as the test set, and use the remaining

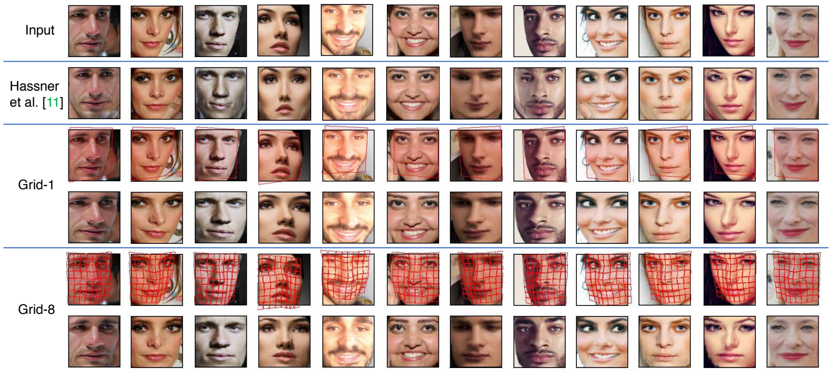


Fig. 4: Qualitative Analysis of SNFace Testset. We sample the data from the SNFace test set with pose, expression, and illumination variations, and visualize the rectified results under different rectification methods.

ones as the training set. The 5-point facial landmarks are detected and the face is aligned with similarity transformation.

Network Architecture. In all the experiments in this paper, we use the GoogLeNet [33, 28] for our recognition network. The rectification network is based on a modified Inception module, which contains fewer parameters and a simpler structure. The rectification network takes very limited additional parameters and time computation compared with the recognition network. The Denoising Autoencoder is designed with a convolutional autoencoder structure. The network details are described in Tab. 1.

Implementation Details. The dimension of the original and rectified face of the rectification network are 128×128, and the pixel level activations are normalized by dividing 255. The Denoising Autoencoder is trained on a subset of the SNFace dataset, which contains 100K faces in canonical views. An end-to-end optimization is conducted after the Denoising Autoencoder is ready. In the training phase, each mini-batch contains 1024 image triplets. We set an equal learning rate for all trainable layers to 0.1, which is shrunk by a factor of 10 once the validation error stops decreasing. The hyper parameters are determined by the validation set as $\lambda_{\text{reg}} = 10.0, \alpha = 0.3$, and $\lambda_{\text{de}} = 1.0$. In all the experiments, we use the same metric learning method with triplet loss. No other data processing and training protocol are used. In the testing phase, we use the Euclidean distance as the verification metric between two face representations.

4.2 What is Learned in Face Rectification?

In this section, we study what is learned in the rectification network. All approaches are evaluated on the SNFace test dataset. We evaluate our model with $n = 8$ (i.e., 64 cells in local homography transformations), referred to as *Grid-8*. We compare with several alternative approaches: the *baseline* model does

Evaluation on SNFace Testset			
Method↓	FAR=10^{-2}	FAR=10^{-3}	FAR=10^{-4}
baseline	92.94	81.76	63.41
baseline-3D	94.02	80.36	58.20
Grid-1	93.49	83.94	66.15
Grid-2	94.02	85.24	68.70
Grid-4	94.38	86.23	71.09
Grid-8\reg	94.10	85.44	69.05
Grid-8	**94.92**	**87.81**	**72.71**

Table 2: Quantitative Results on the SNFace Testset. We compare our method *Grid-8* against several other approaches and report verification accuracy on the SNFace test set.

not have face rectification; the global model *Grid-1* performs the face rectification with global homography transformation; no face prior regularization model *Grid-8\reg* does not have the regularization during training.

Moreover, in order to compare with the 3D face frontalization technique used in face recognition (e.g., 3D alignment used in DeepFace [34]), we process the full SNFace dataset to synthesize frontal views by using a recent face fronalization method created by Hassner et al. [12], and compare with the model trained on this synthesized data (called *baseline-3D*) to verify the effectiveness of our rectification process and joint optimization.

Qualitative Analysis. Fig. 4 depicts the visualization results of the original images and the corresponding rectified images. Obviously, the global homography transformation *Grid-1* can capture coarse geometric information, such as 2D rotation and translation, which is also reported in previous works [44, 39]. However, due to its limited capacity, *Grid-1* is unable to satisfactorily rectify out-of-plane rotation and local geometric details, and generate with notable distortion (e.g., the big nose in the faces with large pose). Hassner et al. [12] improve further, generating good frontal view faces, but the ghosting effect (most faces under large pose in Fig. 4) and the change of facial shape (e.g., nose in the fourth individual in Fig. 4) may introduce further noise to the recognition system. On the other hand, *Grid-8* can capture rich facial structure details. Local geometric warping is detected and approximated by local homographies. Compared with the original image and results from other approaches, the proposed method *Grid-8* greatly reduces geometric variations and induces better canonical view results.

Quantitative Analysis. We report quantitative results under verification protocol in Tab. 2. *Grid-8* achieves the best performance which outperforms the baseline by a large margin up from 63.4% to 72.7% with False Alarm Rate (FAR) at 10^{-4}. The global transformation *Grid-1* consistently improves the recognition performance compared with the *baseline*. But as we have seen in the visualization results, global transformation is limited to its transformation capacity and thus introduces large distortion for recognition.

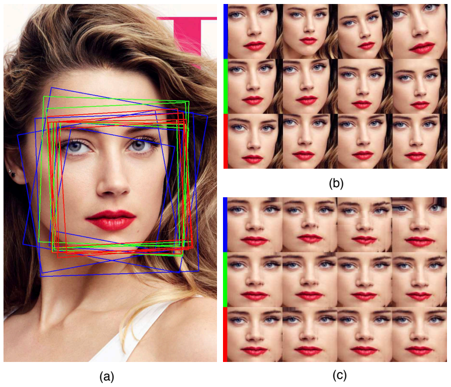


Fig. 5: Synthetic 2D Transformations. Visualization of the image and perturbed samples in the synthetic 2D transformation experiment. (a). Original image, where different color boxes corresponding to different noise levels (red for $\sigma = 0.05d$, green for $\sigma = 0.1d$, and blue for $\sigma = 0.15d$). (b). Cropped faces with noisy landmarks. (c). Rectified faces by our method *Grid-8*. Most of the scale, rotation, and translation variations are reduced.

Table 3: Quantitative Results under Synthetic 2D Transformations. Verification accuracy of our model *Grid-8* and *baseline* at FAR=10^{-2} under 2D transformations with different noise levels.

Method ↓	$\sigma = 0.00d$	$\sigma = 0.05d$	$\sigma = 0.10d$	$\sigma = 0.15d$
baseline	92.94	91.66	86.58	74.95
Grid-8	94.92	93.51	90.35	85.00

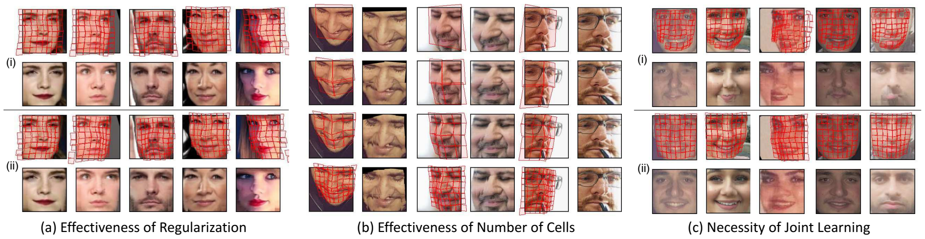
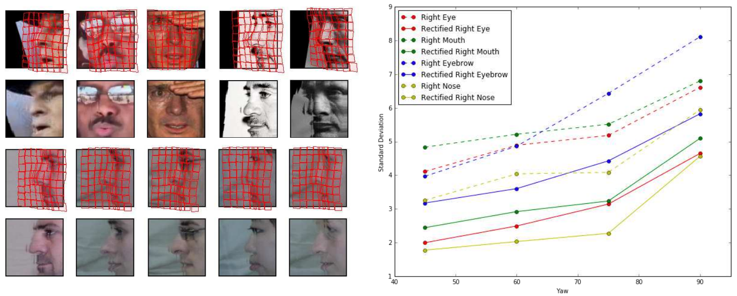

(i)
(ii)
(a) Effectiveness of Regularization
(b) Effectiveness of Number of Cells
(i)
(ii)
(c) Necessity of Joint Learning

Fig. 6: Ablation Studies. (a) i. Rectification without regularization. ii. Rectification with regularization. (b) Rectification with different number of cells. (c) i. Rectification without recognition supervision. ii. Joint learning of rectification and recognition.

The recognition model trained on the synthesized frontal view data *baseline-3D* obtains high performance with FAR at 10^{-2}, better than the *baseline* and *Grid-1* trained on the original data. But the performance drops dramatically, and finally gets 5.2% worse than the *baseline* with FAR at 10^{-4}. On the other hand, our method *Grid-8* consistently outperforms the *baseline-3D* and obtains 14.5% improvement with FAR at 10^{-4}.

Evaluation on Synthetic 2D Transformations. We investigate the effectiveness of face rectification for reducing 2D in-plane transformations, which is typically introduced by facial landmarks. The perturbed data is generated by performing face alignment with noisy landmarks, which are synthesized by adding *i.i.d.* Gaussian noise of variance σ. The Gaussian noise mimics the inaccurate facial landmarks in the real system, and introduces the scale, rotation, and translation variations in the face alignment. We normalize the face size by



(a) Qualitative Results

(b) Facial Landmark Variations

Fig. 7: Evaluation on Challenge Situations. (a). Qualitative results under large pose and occlusion. (b). Comparisons of standard deviation of facial landmarks under different pose variations.

interocular distance d, and generate perturbed data with different noise levels $\sigma = 0.05d, 0.1d, 0.15d$.

Fig. 5 presents the visualization of synthetic data with small (red boxes with $\sigma = 0.05d$), middle (green boxes with $\sigma = 0.1d$) and large (blue boxes with $\sigma = 0.15d$) noise levels. As shown in Fig. 5 (c), the rectification network can generate canonical view faces, which greatly reduce the in-plane variance. Tab. 3 reports the qualitative comparison between the *baseline* and *Grid-8*. We can see that the *baseline* suffers from the large in-plane variations and the accuracy drops rapidly. Meanwhile, the rectification network *Grid-8* yields much better performance even under large geometric variations.

Effectiveness of Regularization. We further explore the effectiveness of regularization. Visualization results of rectified faces are shown in Fig. 6(a). The first two rows present the rectification trained without regularization, and the last two rows show the results with regularization. We can observe that the regularization helps the rectification process generate more canonical view faces, and reduces the cropping variations in the rectified results. Quantitative results are reported in Tab. 2. The regularization achieves 2.4% improvement with FAR at 10^{-3}, and 3.7% improvement with FAR at 10^{-4}.

Number of Partition Cells. We investigate the influence of the number of partition cells in the rectification network. Visualized results of $n = 1, 2, 4, 8$ are presented in Fig. 6(b), and the quantitative results in SNFace test set are shown in Tab. 2. As the number of cells increases, the image distortion introduced in the rectified face decreases, and verification performance increases, benefiting from the local homography transformations.

Necessity of Joint Learning. To evaluate the contribution of joint learning the face rectification and recognition, we introduce an ablation experiment learning each part sequentially. This model first learns the face rectification without the recognition supervision, and then trains the recognition module with the fixed rectification module. Fig. 6(c) provides the qualitative results. The consequences of the lack of recognition supervision is obvious and irreversible. The noisy gradient provided by the Denoising Autoencoder introduces much artifacts and the misalignment objective further drops many face details (e.g. close the mouth in the first and second individuals). On the other hand, joint learning of rectification and recognition can greatly reduce artifacts in the rectification results and keep the most of facial details. The recognition accuracy of this sequential learning model is $91.1\%, 72.3\%, 41.6\%$ with FAR at $10^{-2}, 10^{-3}, 10^{-4}$, which is far below the joint learning model and even the original baseline.

Evaluation on Challenge Situations. Fig. 7(a) presents the rectification results under challenging occlusion situations like large pose and sunglasses. The effects of rectification process is not hallucinating the missing parts. It reduces the geometric variations and does alignment for the visible parts. We further evaluate the variations of facial landmarks on the Multi-PIE dataset [10]. Four facial landmarks in the right side of face are considered and the corresponding standard deviations are calculated. Fig. 7(b) demonstrates the landmark variations in the original and rectified faces under different face pose. Obviously, the variations of each landmark across different poses are much smaller than ones in the original face, which suggests that our rectification process is robust to pose variation and reduce facial geometric variations significantly.

4.3 Evaluation on Public Benchmarks

To verify the cross-data generalization of learned models, we report our performance on four challenge public benchmarks, which cover large pose, expression, and illumination variations. We further report our models trained under the public dataset MS-Celeb-1M [11], referred to as *baseline-Pub* and *Grid-8-Pub*.

LFW and YTF. In the LFW dataset [13], we follow the standard evaluation protocol of unrestricted with labeled outside data, and report the mean accuracy (mAC) of 10-folders verification set. We further follow the identification protocol proposed by Best-Rowden et al. [2], and report the closed-set recognition performance measured by rank-1 rate and the open-set performance measured by Detection and Identification Rate (DIR) with FAR at 1%. In the YTF dataset [38], we follow the standard protocol and report the mAC of 10 folds video verification set. We perform the video-to-video verification by averaging the similarity scores between every pairs of images.

Results on LFW and YTF. Tab. 4 shows our results. In the LFW verification benchmark, our method consistently improves the performance up from 99.05% to 99.68% with the MS-Celeb training set and from 99.15% to 99.70% with the SNFace training set. Our results are comparable with FaceNet [28] but with the considerably smaller training data (10M training faces VS 200M faces). Under the LFW identification protocol, our method boosts the baseline with significant improvements (up from 91.7% to 96.7% in the close-set protocol and from 80.3% to 94.1% in the open-set protocol), and achieves the state-of-the-art. In the YTF benchmark, our method *Grid-8* (95.6%) and *Grid-8-Pub* (95.2%) also provide consistent improvements over the baseline methods *baseline* (94.0%) and *baseline-Pub* (93.4%). Fig. 8 provides the rectification results in LFW and YTF.

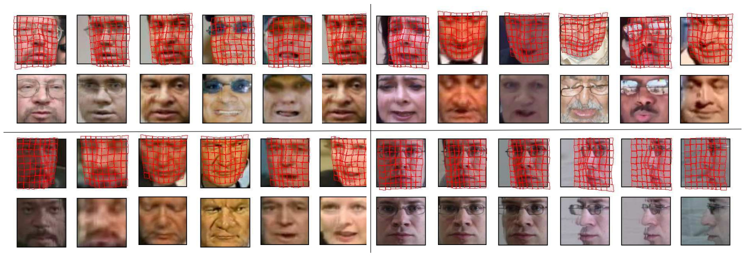


Fig. 8: Qualitative Analysis on Public Benchmarks. Left Top: LFW; Left Bottom: YTF; Right Top: IJB-A; Right Bottom: Multi-PIE.

Multi-PIE. The Multi-PIE dataset [10] contains 754,200 images from 337 subjects, covering large variations in pose and illumination. We follow the protocol from [41], where the last 137 subjects with 13 poses, 20 illuminations and neutral expression are selected for testing. For each subject, we randomly choose one image with frontal pose and neutral illumination as the gallery, and leave all the rest as probe images.

Results on Multi-PIE. Tab. 5 shows our results. Our method outperforms the baseline methods by a large margin, from 44.3% (*baseline-Pub*) to 62.0% (*Grid-8-Pub*) and 65.5% (*baseline*) to 75.4% (*Grid-8*) in the identification rate with face yaw at 90° . We achieve the best performance and outperforms the recent GAN-based face frontalization methods [36, 14, 42]. Moreover, we do not observe the performance degeneration in frontal face, which indicates that our method introduces few artifacts in frontal faces and gains consistent improvement over pose variations. Fig. 8 shows the qualitative results with different pose variations. The rectification process is robust to the change of pose, and reduces the geometric variations for the visible parts.

IJB-A. The IJB-A dataset is a challenge benchmark due to its unconstrained setting. It defines the set-to-set recognition as face template matching. In our evaluation, we do not introduce complicate strategies, and perform the set-to-set recognition via a media pooling followed from the previous method [26]. Specifically, the template feature is extracted by first averaging all image feature with their media ID, and then averaging between medias.

Results on IJB-A. Tab. 6 and Fig. 8 report our results on IJB-A. It is worth pointing out that we employ strong baselines, which achieve 68.5% (*Baseline-Pub*) and 71.3% (*Baseline*) verification accuracy with FAR at 0.001, and 89.6% (*Baseline-Pub*) and 90.4% (*Baseline*) rank-1 identification accuracy. By adding our rectification process, our rectification methods outperform the these strong baselines by a large margin. We achieve 13.8% (*Grid-8-Pub*) and 12.6% (*Grid-8*) improvement on the verification task (with FAR at 0.001), and reduce 25% (*Grid-8-Pub*) and 26% (*Grid-8*) error rate on the rank-1 identification task. It is noteworthy that multiple-frame aggregation methods [40, 18] in the set-to-set recognition scenarios (e.g., IJB-A and YTF) can achieve better performance. These techniques could also apply to our method and is left to the future work.

Table 4: Evaluation on LFW and YTF

Method↓	LFW mAC	LFW Rank-1	LFW DIR@1%	YTF mAC
DeepFace [34]	97.35	64.9	44.5	91.4
VGGFace [24]	99.13	-	-	97.4
FaceNet [28]	99.64	-	-	95.1
DeepID2+ [32]	99.47	95.0	80.7	93.2
WST Fusion [35]	98.37	82.5	61.9	-
SphereFace [18]	99.42	-	-	95.0
RangeLoss [43]	99.52	-	-	93.7
HiReST-9+ [39]	99.03	93.4	80.9	95.4
Baseline-Pub	99.05	88.9	78.8	93.4
Grid-8-Pub	99.68	96.4	93.1	95.2
Baseline	99.15	91.7	80.3	94.0
Grid-8	99.70	96.7	94.1	95.6

Table 5: Evaluation on Multi-PIE

Method↓	0°	15°	30°	45°	60°	75°	90°
Yim et al. [41]	99.5	95.0	88.5	79.9	61.9	-	-
DRGAN [36]	97.0	94.0	90.1	86.2	83.2	-	-
TPGAN [14]	-	98.7	98.1	95.4	87.7	77.4	64.6
FF-GAN [42]	95.7	94.6	92.5	89.7	85.2	77.2	61.2
Baseline-Pub	100.0	100.0	100.0	98.9	92.9	78.4	44.3
Grid-8-Pub	100.0	100.0	100.0	99.3	96.1	86.7	62.0
Baseline	100.0	100.0	100.0	100.0	98.7	92.6	65.5
Grid-8	100.0	100.0	100.0	100.0	99.2	94.7	75.4

Table 6: Evaluation on IJB-A

Method↓	Verification		Identification	
Metric→	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
PAM [20]	73.3 ± 1.8	55.2 ± 3.2	77.1 ± 1.6	88.7 ± 0.9
Masi et al.[21]	88.6 ± 1.7	72.5 ± 4.4	90.6 ± 1.3	96.2 ± 0.7
TripEmbd. [26]	90.0 ± 1.0	81.3 ± 2.0	93.2 ± 1.0	-
TempAdpt. [7]	93.9 ± 1.3	83.6 ± 2.7	92.8 ± 1.0	97.7 ± 0.4
DRGAN [36]	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1
FFGAN [42]	85.2 ± 1.0	66.3 ± 3.3	90.2 ± 0.6	95.4 ± 0.5
Baseline-Pub	86.6 ± 1.8	68.5 ± 3.9	89.6 ± 1.3	95.2 ± 0.5
Grid-8-Pub	91.5 ± 0.8	82.3 ± 1.9	92.2 ± 1.0	96.0 ± 0.5
Baseline	88.7 ± 1.9	71.3 ± 3.9	90.4 ± 1.1	95.4 ± 0.7
Grid-8	92.1 ± 0.8	83.9 ± 1.4	92.9 ± 1.0	96.2 ± 0.5

5 Conclusion

In this paper, we develop a method called *GridFace* to reduce facial geometric variations. We propose a novel non-rigid face rectification method by local homography transformations, and regularize it by imposing natural frontal face distribution with a Denoising Autoencoder. Empirical results show our method greatly reduces geometric variations and improves the recognition performance.

References

1. Alain, G., Bengio, Y.: What regularized auto-encoders learn from the data-generating distribution. The Journal of Machine Learning Research **15**(1), 3563–3593 (2014)
2. Best-Rowden, L., Han, H., Otto, C., Klare, B.F., Jain, A.K.: Unconstrained face recognition: Identifying a person of interest from a media collection. IEEE Transactions on Information Forensics and Security **9**(12), 2144–2157 (2014)
3. Bhattarai, B., Sharma, G., Jurie, F.: Cp-mtml: Coupled projection multi-task metric learning for large scale face retrieval. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
4. Bigdeli, S.A., Jin, M., Favaro, P., Zwicker, M.: Deep mean-shift priors for image restoration. arXiv preprint arXiv:1709.03749 (2017)
5. Chen, D., Hua, G., Wen, F., Sun, J.: Supervised transformer network for efficient face detection. In: European Conference on Computer Vision. pp. 122–138. Springer (2016)
6. Cole, F., Belanger, D., Krishnan, D., Sarna, A., Mosseri, I., Freeman, W.T.: Synthesizing normalized faces from facial identity features. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
7. Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., Zisserman, A.: Template adaptation for face verification and identification. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on. pp. 1–8. IEEE (2017)
8. Du, L., Ling, H.: Cross-age face verification by coordinating with cross-face age verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
10. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)
11. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016)
12. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
13. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
14. Huang, R., Zhang, S., Li, T., He, R.: Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in Neural Information Processing Systems 28. pp. 2017–2025 (2015)
16. Kanazawa, A., Jacobs, D.W., Chandraker, M.: Warpnet: Weakly supervised matching for single-view reconstruction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
17. Lin, C.H., Lucey, S.: Inverse compositional spatial transformer networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)

18. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
19. Lu, B.L., Zheng, J., Chen, J.C., Chellappa, R.: Pose-robust face verification by exploiting competing tasks. In: Applications of Computer Vision (WACV) (June 2017)
20. Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
21. Masi, I., Tran, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do We Really Need to Collect Millions of Faces for Effective Face Recognition?, pp. 579–596 (2016)
22. Meinhardt, T., Moller, M., Hazirbas, C., Cremers, D.: Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
23. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
24. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
25. Peng, X., Yu, X., Sohn, K., Metaxas, D.N., Chandraker, M.: Reconstruction-based disentanglement for pose-invariant face recognition. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
26. Sankaranarayanan, S., Alavi, A., Castillo, C.D., Chellappa, R.: Triplet probabilistic embedding for face verification and clustering. In: Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on. pp. 1–8. IEEE (2016)
27. Särelä, J., Valpola, H.: Denoising source separation. Journal of machine learning research **6**(Mar), 233–272 (2005)
28. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
29. Sønderby, C.K., Caballero, J., Theis, L., Shi, W., Huszár, F.: Amortised map inference for image super-resolution. arXiv preprint arXiv:1610.04490 (2016)
30. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems 27, pp. 1988–1996 (2014)
31. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)
32. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
33. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
34. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2014)

35. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Web-scale training for face identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
36. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
37. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision. pp. 499–515. Springer (2016)
38. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 529–534 (2011)
39. Wu, W., Kan, M., Liu, X., Yang, Y., Shan, S., Chen, X.: Recursive spatial transformer (rest) for alignment-free face recognition. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
40. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
41. Yim, J., Jung, H., Yoo, B., Choi, C., Park, D., Kim, J.: Rotating your face using multi-task deep neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
42. Yin, X., Yu, X., Sohn, K., Liu, X., Chandraker, M.: Towards large-pose face frontalization in the wild. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
43. Zhang, X., Fang, Z., Wen, Y., Li, Z., Qiao, Y.: Range loss for deep face recognition with long-tailed training data. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
44. Zhong, Y., Chen, J., Huang, B.: Toward end-to-end face recognition through alignment learning. IEEE Signal Processing Letters **24**(8), 1213–1217 (Aug 2017). https://doi.org/10.1109/LSP.2017.2715076
45. Zhu, X., Lei, Z., Yan, J., Yi, D., Li, S.Z.: High-fidelity pose and expression normalization for face recognition in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
46. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity-preserving face space. In: The IEEE International Conference on Computer Vision (ICCV) (December 2013)
47. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. In: Advances in Neural Information Processing Systems 27. pp. 217–225 (2014)