# Face Synthesis for Eyeglass-Robust Face Recognition

Jianzhu Guo, Xiangyu Zhu*, Zhen Lei and Stan Z. Li

CBSR&NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
{jianzhu.guo,xiangyu.zhu,zlei,szli}@nlpr.ia.ac.cn

**Abstract.** In the application of face recognition, eyeglasses could significantly degrade the recognition accuracy. A feasible method is to collect large-scale face images with eyeglasses for training deep learning methods. However, it is difficult to collect the images with and without glasses of the same identity, so that it is difficult to optimize the intra-variations caused by eyeglasses. In this paper, we propose to address this problem in a virtual synthesis manner. The high-fidelity face images with eyeglasses are synthesized based on 3D face model and 3D eyeglasses. Models based on deep learning methods are then trained on the synthesized eyeglass face dataset, achieving better performance than previous ones. Experiments on the real face database validate the effectiveness of our synthesized data for improving eyeglass face recognition performance.

**Keywords:** Face recognition, 3D eyeglass fitting, face image synthesis

## 1 Introduction

In recent years, deep learning based face recognition systems [1,2,3,4] have achieved great success, such as Labeled Faces in the Wild (LFW) [5], YouTube Faces DB (YFD) [6], and MegaFace [7].

However, in practical applications, there are still extra factors affecting the face recognition performance, e.g., facial expression, poses, occlusions etc. Eyeglasses, especially black-framed eyeglasses significantly degrade the face recognition accuracy (see Table 4). There are three common categories of eyeglasses: thin eyeglasses, thick eyeglasses, and sunglasses. In this work, we mainly focus on the category of thick black-framed eyeglasses, since the effects of thin eyeglasses are tiny, while the impact of sunglasses are too high because of serious identity information loss in face texture.

The main contributions of this work include: 1) A eyeglass face dataset named MeGlass, including about $1.7K$ identities, is collected and cleaned for eyeglass face recognition evaluation. It will be made public on https://github.com/cleardusk/MeGlass. 2) A virtual eyeglass face image synthesis method is proposed. An eyeglass face training database named MsCeleb-Eyeglass is generated, which helps improve the robustness to eyeglass. 3) A novel metric learning

* Corresponding author

method is proposed to further improves the face recognition performance, which is designed to adequately utilize the synthetic training data.

The rest of this paper is organized as follows. Section 2 reviews several related works. Our proposed methods are described in Section 3. The dataset description is in Section 4. Extensive experiments are conducted in Section 5 to validate the effectiveness of our synthetic training data and loss function. Section 6 summarizes this paper.

## 2 Related Work

**Automatic eyeglasses removal.** Eyeglasses removal is another method to reduce the effect of eyeglasses on face recognition accuracy. Several previous works [9,10,11,12] have studied on automatic eyeglasses removal. Saito et al. [9] constructed a non-eyeglasses PCA subspace using a group of face images without eyeglasses, one new face image was then projected on it to remove eyeglasses. Chenyu Wu et al. [10] proposed an intelligent image editing and face synthesis system for automatic eyeglasses removal, in which eyeglasses region was first detected and localized, then the corrupted region was synthesized adopting a statistical analysis and synthesis approach. Park et al. [12] proposed a recursive process of PCA reconstruction and error compensation to further eliminate the traces of thick eyeglasses. However, these works did not study the quantitative effects of eyeglasses removal on face recognition performance.

**Virtual try-on.** Eyeglass face image synthesis is similar to virtual eyeglass try-on. Recently, eyeglasses try-on has drawn attentions in academic community. Niswar et al. [15] first reconstructed 3D head model from single image, 3D eyeglasses were next fitted on it, but it lacked the rendering and blending process compared with our synthesis method. Yuan X et al. [13] proposed a interactive real time virtual 3D eyeglasses try-on system. Zhang, Q et al. [14] firstly took the refraction effect of corrective lenses into consideration. They presented a system for trying on prescription eyeglasses, which could produce a more real look of wearing eyeglasses.

**Synthetic images for training.** Recently, synthetic images generated from 3D models have been studied in computer vision [17,18,19,20]. These works adopted 3D models to render images for training object detectors and viewpoint classifiers. Because of the limited number of 3D models, they tweaked the rendering parameters to generate more synthetic samples to maximize the model usage.

## 3 Proposed Method

### 3.1 Eyeglass image synthesis

We describe the details of eyeglass face synthesis in this section. To generate faces with eyeglasses, we estimate the positions of the 3D eyeglasses based on the fitted 3D face model and then render the 3D eyeglasses on the original face

Fig. 1: Four pairs of origin-synthesis images selected from MsCeleb.

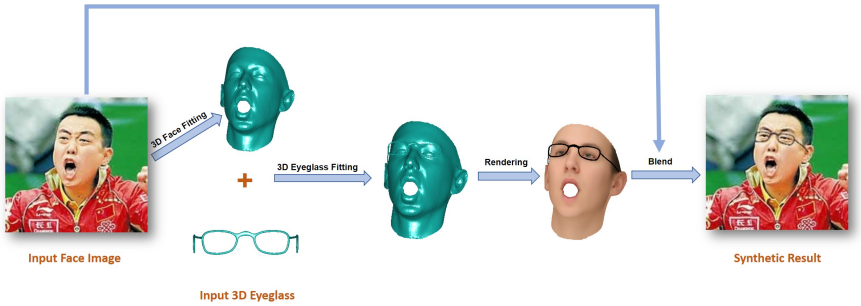


Fig. 2: The pipeline of eyeglass synthesis.

images. The whole pipeline of our eyeglass faces synthesis is shown in Fig. 2. Firstly, we reconstruct the 3D face model based on pose adaptive 3DMM fitting method [22], which is robust to pose. Secondly, the 3D eyeglass is fitted on the reconstructed 3D face model. The fitting is based on the corresponding anchor points on the 3D eyeglass and 3D fitted face model, where the indices of these anchor points are annotated beforehand. Then z-buffer algorithm and Phong illumination model are adopted for rendering, and the rendered eyeglass image is blended on the original image to generate the final synthetic result.

The 3D eyeglass fitting problem is formed as Eq. 1, where $f$ is the scale factor, $Pr$ is the orthographic projection matrix, $p_g$ is the anchor points on 3D eyeglass, $p_f$ is the anchor points on reconstructed 3D face model, $R$ is the $3 \times 3$ rotation matrix determined by pitch($\alpha$), yaw($\beta$), and roll($\gamma$) and $t_{3d}$ is the translation vector.

$$\arg \min_{f,Pr,R,t_{3d}} ||f * Pr * R * (p_g + t_{3d}) - p_f||, \tag{1}$$

Although the amount of images of MsCeleb is large, the model may overfit during training if the patterns of synthetic eyeglasses are simple. To increase the diversity of our synthetic eyeglass face images, we inject randomness into two steps of our pipeline: 3D eyeglass preparation and rendering. For 3D eyeglasses, we prepare four kinds of eyeglasses with different shapes and randomly select one as input. For eyeglass rendering, we explore three sets of parameters: light condition, pitch angle and vertical transition of eyeglass. For the light condition, the energies and directions are randomly sampled. Furthermore, to simulate the real situations of eyeglass wearing, we add small perturbations to the pitch angle ($[-1.5, 0.8]$) and vertical transition ($[1, 2]$ pixel). Finally, we put together the synthetic eyeglass face images with original images as our training datasets.

Table 1: Our ResNet-22 network structure. Conv3.x, Conv4.x and Conv5.x indicates convolution units which may contain multiple convolution layers and residual blocks are shown in double-column brackets. E.g., $[3 \times 3, 128] \times 3$ denotes 3 cascaded convolution layers 128 feature maps with filters of size $3 \times 3$, and S2 denotes stride 2. The last layer is global pooling.

| **Layers** | **22-layer CNN** |
|---|---|
| Conv1.x | [5×5, 32]×1, S2 |
| Conv2.x | [3×3, 64]×1, S1 |
| Conv3.x | $\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 3$, S2 |
| Conv4.x | $\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 226 \end{bmatrix} \times 4$, S2 |
| Conv5.x | $\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$, S2 |
| Global Pooling | 512 |

### 3.2 Network and loss

**Network** We adapt a 22 layers residual network architecture based on [24] to fit our task. The original ResNet is designed for ImageNet [25], the input image size is $224 \times 224$, while ours is $120 \times 120$. Therefore, we substitute the original $7 \times 7$ convolution in first layer with $5 \times 5$ and stack one $3 \times 3$ convolution layer to preserve dimensions of feature maps. The details of our ResNet-22 are summarized in Table 1.

**Loss** Due to the disturbance of eyeglass on feature discrimination, we propose the Mining-Contrasive loss based on [26] to further enlarge the inter-identity differences and reduce intra-identity variations. The form of our proposed loss is in Eq. 2.

$$L_{mc} = -\frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} d(f_i, f_j) + \frac{1}{2|\mathcal{N}|} \sum_{(i,j)\in\mathcal{N}} d(f_i, f_j). \tag{2}$$

Where $f_i$ and $f_j$ are vectors extracted from two input image samples, $\mathcal{P}$ is hard positive samples set, $\mathcal{N}$ is hard negative samples set, $d(f_i, f_j) = \frac{f_i \cdot f_j}{||f_i||_2 ||f_j||_2}$ is cosine similarity between extracted vectors.

**Gradual sampling**. Besides, we employ the gradual process into data sampling to make the model fit the synthetic training images in a gentle manner. In naive sampling, the probability of eyeglass face image of each identity is fixed at 0.5. It means that we just brutally mix MsCeleb and MsCeleb-Eyeglass datasets. We then generalize the sampling probability as $p = \lambda \cdot n + p_0$, where $n$ is the

Table 2: Summary of dataset description. G and NG indicate eyeglass and non-eyeglass respectively. Mixture means the MsCeleb and MsCeleb-Eyeglass.

| **Dataset** | **Identity** | **Images** | **G** | **NG** |
|---|---|---|---|---|
| MeGlass | 1,710 | 47,917 | 14,832 | 33,085 |
| Testing set | 1,710 | 6,840 | 3,420 | 3,420 |
| Training set(MsCeleb) | 78,765 | 5,001,877 | - | - |
| Training set(Mixture) | 78,765 | 10,003,754 | - | - |

Fig. 3: Sample images of our testing set. For each identity, we show two faces with and without eyeglasses.

number of iterations, $\lambda$ is the slope coefficient determining the gradual process, $p_0$ is the initialized probability value.

## 4 Dataset Description

In this section, we describe our dataset in detail and the summary is shown in Table 2.

### 4.1 Testing set

We select real face images with eyeglass from MegaFace [7] to form the MeGlass dataset. We first apply an attribute classifier to classify the eyeglass and non-eyeglass face images automatically. After that, we select the required face images manually from the attibute-labeled face images. Our MeGlass dataset contains 14,832 face images with eyeglasses and 33,087 images without eyeglasses, from 1,710 subjects.

To be consistent with the evaluation protocol (in Section 4.3), we select two faces with eyeglasses and two faces without eyeglasses from each identity to build our testing set and the total number of images is 6,880. Fig. 3 shows some examples of testing set with and without eyeglasses.

### 4.2 Training set

Two types of training set are adopted, one is only the MsCeleb and the other is the mixture of MsCeleb with synthetic MsCeleb-Eyeglass. Our MsCeleb clean list has $78,765$ identities and $5,001,877$ images, which is slightly modified from [23]. For each image, we synthesize a eyeglass face image using the method proposed in Section 3.1. Therefore, there are totally $10,003,754$ images from $78,765$ subjects in the mixture training set.

### 4.3 Evaluation protocol

In order to examine the effect of eyeglass on face recognition thoroughly, we propose four testing protocols to evaluate different methods.

I) All gallery and probe images are without eyeglasses. There are two non-eyeglass face images per person in gallery and probe sets, respectively.
II) All gallery and probe images are with eyeglasses. There are two eyeglasses face images per person in gallery and probe sets, respectively.
III) All gallery images are without eyeglasses, and all probe images are with eyeglasses. There are two non-eyeglass face images per person in gallery set and there are two eyeglass face images per person in probe set.
IV) Gallery images contain both eyeglass images and non-eyeglass images, so as probe images. There are four face images (including two non-eyeglass and two eyeglass face images) per person for gallery and probe sets.

## 5 Experiments

Firstly, we evaluate the impact of eyeglasses on face recognition. Second, several experiments are conducted to study the effect of synthetic training data and proposed loss. In experiments, two losses including the classification loss A-Softmax and the metric learning based contrastive loss are investigated. Totally there are four deep learning models are trained based on different losses and training sets. Table 3 lists the four deep face models. For ResNet-22-A, we apply A-Softmax loss to learn the model from original MsCeleb dataset. We then finetune the model on MsCeleb dataset using contrastive loss to obtain ResNet-22-B. We also finetune the ResNet-22-A model on MsCeleb and its synthetic eyeglasses database to obtain ResNet-22-C. The ResNet-22-D is finetuned from base model ResNet-22-A using gradual sampling strategy with the slope coefficient $\lambda$ of 0.00001 and $p_0$ of 0.

### 5.1 Experiments Settings

Our experiments are based Caffe [27] framework and Tesla M40 GPU. All face images are resized to size $120 \times 120$, then being normalized by subtracting 127.5 and being divided by 128. We use SGD with a mini-batch size of 128 to optimize the network, with the weight decay of 0.0005 and momentum of 0.9. Based on these configurations, the training speed can reach about 260 images per second on single GPU and the inference speed is about 1.5ms per face image.

Table 3: The configuration settings of different models. ResNet-22-B, ResNet-22-C and ResNet-22-D are all finetuned from ResNet-22-A. GS indicates gradual sampling.

| **Model** | **Training Data** | **Loss** | **Strategy** |
|---|---|---|---|
| ResNet-22-A | MsCeleb | A-Softmax [4] | - |
| ResNet-22-B | MsCeleb | Mining-Contrasive | Finetune |
| ResNet-22-C | Mixture | Mining-Contrasive | Finetune |
| ResNet-22-D | Mixture | Mining-Contrasive | Finetune+GS |

Table 4: Recognition performance (%) of ResNet-22-A following protocols I-IV.

| Protocol | TPR@FAR=$10^{-4}$ | TPR@FAR=$10^{-5}$ | TPR@FAR=$10^{-6}$ | Rank1 |
|---|---|---|---|---|
| I | 96.14 | 91.49 | 84.68 | 98.48 |
| II | 94.09 | 86.55 | 69.21 | 96.90 |
| III | 88.13 | 74.72 | 59.34 | 95.61 |
| IV | 78.17 | 60.25 | 41.36 | 92.31 |

### 5.2 Effect of eyeglass on face recognition

In this experiment, we use the original MsCeleb database only as the training set to examine the robustness of traditional deep learning model to eyeglasses.

Table 4 shows the results of ResNet-A model tested on four protocols. From the results, one can see that the ResNet-A model achieves high recognition accuracy on protocol I, which is without eyeglass occlusion. However, its performance degrades significantly on protocols II-IV, where eyeglasses occlusion occurs in gallery or probe set, especially for the TPR at low FAR. It indicates that the performance of deep learning model is sensitive to eyeglasses occlusion.

### 5.3 Effectiveness of synthetic data and proposed loss

Table 5: Recognition performance (%) of ResNet-22-B following protocols I-IV.

| Protocol | TPR@FAR=$10^{-4}$ | TPR@FAR=$10^{-5}$ | TPR@FAR=$10^{-6}$ | Rank1 |
|---|---|---|---|---|
| I | 96.61 | 91.26 | 87.02 | 98.60 |
| II | 94.91 | 87.87 | 73.27 | 97.08 |
| III | 89.55 | 76.86 | 62.56 | 96.02 |
| IV | 81.96 | 65.68 | 46.71 | 94.18 |

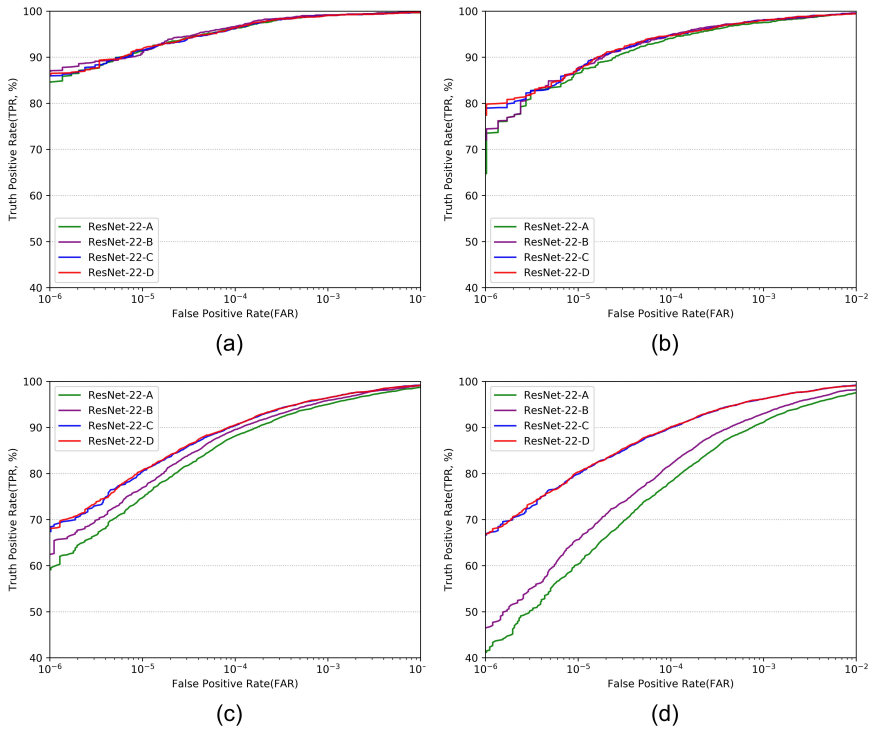


Fig. 4: From (a) to (d): ROC curves of protocols I-IV. For protocols I-II, the curves are almost the same. While for protocols III-IV, ResNet-22-C and ResNet-D models outperform the other two. Especially in protocol IV, they outperform by a large margin (better view on electronic version).

For comparison, we further train deep face model, ResNet-C, using the mixture of the original MsCeleb and its synthesized eyeglasses version MsCeleb-Eyeglass. Table 5 and Table 6 show the comparison results of ResNet-B and ResNet-C following four protocols. It can be seen that using our synthesized eyeglass face images, it significantly improves the face recognition performance following protocol III-IV, especially at low FAR. It enhances about 20 percent when FAR=$10^{-6}$ on protocol IV, which is the hardest case in four configurations, indicating the effectiveness of virtual face synthesis data for the robustness improvement of face deep model. Moreover, with the face synthesis data , the proposed loss function with gradual sampling, model ResNet-22-D achieves the best results on four protocols.

Finally, we also plot the ROC curves for four protocols in Fig. 4 to further validate the effectiveness of our synthetic training dataset and proposed loss function.

Table 6: Recognition performance (%) of ResNet-22-C following protocols I-IV.

| Protocol | TPR@FAR=$10^{-4}$ | TPR@FAR=$10^{-5}$ | TPR@FAR=$10^{-6}$ | Rank1 |
|---|---|---|---|---|
| I | 96.20 | 91.58 | 85.94 | 98.19 |
| II | 94.80 | 87.31 | 78.89 | 96.73 |
| III | 90.35 | 80.40 | 67.93 | 96.67 |
| IV | 89.94 | 79.88 | 66.82 | 96.67 |

Table 7: Recognition performance (%) of ResNet-22-D following protocols I-IV.

| Protocol | TPR@FAR=$10^{-4}$ | TPR@FAR=$10^{-5}$ | TPR@FAR=$10^{-6}$ | Rank1 |
|---|---|---|---|---|
| I | 96.37 | 91.99 | 86.37 | 98.30 |
| II | 94.68 | 87.54 | 78.68 | 96.78 |
| III | 90.54 | 80.71 | 68.10 | 96.75 |
| IV | 90.14 | 80.32 | 66.92 | 96.73 |

## 6 Conclusion

In this paper, we propose a novel framework to improve the robustness of face recognition with eyeglasses. We synthesize face images with eyeglasses as training data based on 3D face reconstruction and propose a novel loss function to address this eyeglass robustness problem. Experiment results demonstrate that our proposed framework is rather effective. In future works, the virtual-synthesis method may be extended to alleviate the impact of other factors on the robustness of face recognition encountered in real life application.

## 7 Acknowledgments

This work was supported by the Chinese National Natural Science Foundation Projects #61473291, #61572536, #61572501, #61573356, the National Key Research and Development Plan (Grant No.2016YFC0801002), and AuthenMetric R&D Funds.

## References

1. Taigman Y, Yang M, Ranzato MA, Wolf L: Deepface: Closing the gap to human-level performance in face verification. CVPR (2014)
2. Sun Y, Wang X, Tang X: Deep learning face representation from predicting 10,000 classes. CVPR (2014)
3. Schroff F, Kalenichenko D, Philbin J: Facenet: Facenet: A unified embedding for face recognition and clustering. CVPR (2015)
4. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L: Sphereface: Deep hypersphere embedding for face recognition. CVPR (2017)

5. Huang GB, Ramesh M, Berg T, Learned-Miller E: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report (2007)
6. Wolf L, Hassner T, Maoz I: Face recognition in unconstrained videos with matched background similarity. CVPR (2011)
7. Kemelmacher-Shlizerman I, Seitz SM, Miller D, Brossard E: The megaface benchmark: 1 million faces for recognition at scale. CVPR (2016)
8. Guo Y, Zhang L, Hu Y, He X, Gao J: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. ECCV (2016)
9. Saito Y, Kenmochi Y, Kotani K: Estimation of eyeglassless facial images using principal component analysis. ICIP (1999)
10. Wu C, Liu C, Shum HY, Xy YQ, Zhang Z: Automatic eyeglasses removal from face images. TPAMI (2004)
11. Du C, Su G: Eyeglasses removal from facial images. PR (2005)
12. Park JS, Oh YH, Ahn SC, Lee SW: Glasses removal from facial image using recursive error compensation. TPAMI (2005)
13. Yuan X, Tang D, Liu Y, Ling Q, Fang L: From 2D to 3D. TCSVT (2017)
14. Zhang Q, Guo Y, Laffont PY, Martin T, Gross M: A Virtual Try-On System for Prescription Eyeglasses. IEEE COMPUT GRAPH (2017)
15. Niswar A, Khan IR, Farbiz F: Virtual try-on of eyeglasses using 3D model of the head. VRCAI (2011)
16. Chen W, Wang H, Li Y, Su H, Wang Z, Tu C, Lischinski D, Cohen-Or D, Chen B: Synthesizing training images for boosting human 3d pose estimation. 3DV (2016)
17. Su H, Qi CR, Li Y, Guibas LJ: Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. ICCV (2015)
18. Massa F, Russell BC, Aubry M: Deep exemplar 2d-3d detection by adapting from real to rendered views. CVPR (2016)
19. Stark M, Goesele M, Schiele B: Back to the Future: Learning Shape Models from 3D CAD Data. BMVC (2010)
20. Liebelt J, Schmid C: Multi-view object class detection with a 3d geometric model. CVPR (2010)
21. Yi D, Li SZ: Learning sparse feature for eyeglasses problem in face recognition. FG (2011)
22. Zhu X, Lei Z, Yan J, Yi D, Li SZ: High-fidelity pose and expression normalization for face recognition in the wild. CVPR (2016)
23. Wu X, He R, Sun Z, Tan T: A Light CNN for Deep Face Representation with Noisy Labels. arXiv preprint arXiv:1511.02683 (2015)
24. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. CVPR (2016)
25. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L: Imagenet: A large-scale hierarchical image database. CVPR (2009)
26. Sun Y, Chen Y, Wang X, Tang X: Deep learning face representation by joint identification-verification. NIPS (2014)
27. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. ACM Multimedia (2014)