Cross-Pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments

Tianyue Zheng, Weihong Deng
Beijing University of Posts and Telecommunications
Beijing 100876,China
2231135739@qq.com, whdeng@bupt.edu.cn

Abstract

Labeled Faces in the Wild (LFW) database has been widely utilized as the benchmark of unconstrained face verification. Recently, due to big data driven machine learning methods, the performance on the database approaches nearly 100%. However, we argue that this accuracy may be too optimistic. Besides illuminations, occlusions and expressions which have been considered as intra-class variations in LFW positive pairs, cross-pose faces of the same individual is another challenge in face recognition. Therefore, we construct a Cross-Pose LFW (CPLFW) database to add pose influence in face recognition. Also, considering that in the psychology literature, the different-identity pairs should show people of the same gender and race, negative pairs in CPLFW are selected using people with the same gender and race. This assures that performance on the database indicates true ability to distinguish individuals using their identities instead of depending on differences in gender and race. We evaluate some deep learning methods on the new database. Compared to the accuracy on LFW, the accuracy drops about 12%-20% on CPLFW.

1. Introduction

Face verification attempts to verify whether the given two face images represent the same person or two different people. It is often assumed that neither of the photos in testing set contains a person appeared in previous training set.

Labeled Faces in the Wild (LFW) database [6] has been widely used as benchmark to study face verification. To form the dataset, Huang et al. used photos collected as part of the Berkeley Faces in the Wild project [1, 2]. These photos were captured in uncontrolled environments with a wide variety of settings, expressions and lightning. In [6], Huang et al. manually cleaned the data, designed new protocols and released the dataset named 'Labeled Faces in the Wild'. The database includes 13,233 face images of 5,749 individuals. For comparison purpose, the dataset were separated into 10 non-repeating subsets of image pairs for cross validation. Each subset contains 300 positive pairs (two images from the same person) and 300 negative pairs (two images from different people). When the database is used only for testing, all the pairs are used to obtain the performance results. Since then, many researches have been made to obtain better performance upon this database.

While many deep learning methods have reached nearly saturated accuracy on the standard Labeled Faces in the Wild, researchers only solve part of the face verification problem in real world situation. Through inspecting LFW database, one can find two limiting factors which can be improved to better simulate real world condition. One is that the pose difference of a positive pair is not big enough. Although positive pairs in LFW have different light and expressions, most images are near frontal. The lack of pose difference in intra-class variations can be a problem.

Also, according to the research in [10], the imposter distribution should be approximately designed to measure face identification. If we intend to determine a benchmark for face verification, discrimination of the positive pairs and negative pairs should be based on differences in identity only. For positive pairs, it is clear that both faces are of the same gender, race. For the randomly selected negative pairs in LFW, this is generally not true. This results in many face pairs which are trivial to distinguish (e.g. a male and a female). The different attribute distribution between positive pairs and negative pairs should arose enough attention.

In this paper, we consider breaking the two limitation factors of LFW. We reinvent the LFW database by two steps. First, we search images with large pose variations using identities in LFW to form positive pairs. Second, we select negative pairs using individuals with the same race and gender so that non-matched identity pairs differ only in identity. The new database, called Cross-Pose LFW (CPLFW) is collected by crowd-sourcing efforts. The database can be viewed and downloaded at the follow-

Figure 1. Comparison of the same individual in LFW and CPLFW. Pose difference in CPLFW is more obvious.

ing web address: `http://www.whdeng.cn/CPLFW/index.html`. Comparison of the same person pictures in LFW and CPLFW is shown in Figure 1 and according to the pictures we can see that pose difference in CPLFW is more obvious.

We name the new dataset Cross-Pose LFW, the prefix "Cross-Pose" suggests that pose variations of the same individual has been considered as a crucial intra-class variation which better simulates real world face verification situation. Though the images are different from those in LFW, we use the same identities of each fold in LFW and maintain the verification protocols which means our database is a extension of LFW, so the name of our database still includes LFW. There are three motivations behind the construction of CPLFW benchmark as follows:

- Establishing a relatively more difficult database to evaluate the performance of real world face verification so the effectiveness of several face verification methods can be fully justified.

- Continuing the intensive research on LFW with more realistic consideration on pose intra-class variation and fostering the research on cross-pose face verification in unconstrained situation. The challenge of CPLFW emphasizes pose difference to further enlarge intra-class variance. Also, negative pairs are deliberately selected to avoid different gender or race. CPLFW considers both the large intra-class variance and the tiny inter-class variance simultaneously.

- Maintaining the data size, the face verification protocol which provides a 'same/different' benchmark and the same identities in LFW, so one can easily apply CPLFW to evaluate the performance of face verification.

2. Related Works

Face recognition is a popular problem in computer vision for many reasons. First, it is easy to formulate well-posed problem and collect data since individuals come with their name labels. Second, it is worth studying because it is a protruding example of fine-grained classification. Third, face recognition problem is of great importance and can be applied to wide ranges of scenarios. For all these reasons, face recognition has become an area which is popular in the vision community.

Typically, there are two types of tasks for face recognition. One is face identification which means that given gallery set and query set, for a given image in the query set, we want to find the most similar face in gallery set and use the identity of the similar face as the identity of the query image. The other is face verification which determines whether two given images belong to the same person.

Early face datasets were almost collected under controlled environments such as PIE [14], FERET [12] and a very high performance can be obtained on these constrained datasets. However, most models learned from these datasets do not work well in practical applications due to the complexity of faces in real world situation. To improve the generalization of face recognition methods, the interests of datasets gradually changed from controlled environment to uncontrolled environment. And so a milestone dataset Labeled Faces in the Wild (LFW) [6] was established in 2007. Compared to the benchmark dataset before, the biggest difference of LFW is that the images were obtained from Internet rather than acquired under several pre-defined environments. Due to the uncontrolled environment, LFW has various illuminations, expressions, resolutions and these factors are gathered in random way.

Recently, several new face recognition database has been collected to study face recognition and verification. These included CASIA database [16], Megaface [8], IJB-A [7] and FaceScrub [9]. The CASIA dataset [16] consists of 494,414 images of 10,575 subjects. The FaceScrub dataset [9] contains 106,863 images of 530 celebrities collected from the web. Each person has an average number of nearly 200 images. Though the percentage of correct labels is difficult to know, these large and deep databases are useful for

researchers to train face recognition system with complex framework.

Except for the databases used for training, new protocols and benchmarks have also been proposed for face recognition problem. MegaFace [8] was designed to study large scale face recognition. The goal of this dataset is to evaluate the performance of current face recognition algorithms with up to a million distractors. Images were derived from the Yahoo 100 Million Flicker creative commons data set [13]. All of the images in Megaface were first registered in a gallery with one image each person. Then for each subject in FaceScrub [9], one image was used in the gallery and the rest of the images of the person were used as testing images in an identification paradigm. So the goal of face recognition task was to identify the only one matching images in the 1,000,001 individuals. The dataset was established because many applications require accurate identification at plenty scale. It emphasises the ability to identify individuals in very large galleries, or in the open set recognition problem. IJB-A dataset [7] was introduced to push the frontiers of unconstrained face detection and recognition. The database contains 500 individuals with manually localized face images. It is a mix of images and videos which contain full pose variation and can be joint used for face recognition and face detection. The dataset supports both face recognition and face verification.

Many datasets have been designed to measure the performance using criteria that are more strict than that of LFW. For verification, the verification rates at 0.1% false acceptance rate. For identification, rank-1 recognition accuracy on a gallery of millions of people is designed. These protocols and datasets may also involve many comparisons between different poses of the same person. However, the pose difference occurs due to large amount data of the same person, rather than human operation. In addition, these new databases lose the feature of LFW as the easy-to-use, low barriers to entry. In contrast, we manually add pose difference to the same person to enlarge intra-class variations while at the same time using the people with same race and gender as negative pairs to avoid attribute difference influence of positive pairs and negative pairs in face verification. Meanwhile, we design the database by strictly following the protocols of LFW so that researchers need not to do any changes when using the new dataset. These characteristics make the proposed CPLFW database totally different from those datasets above.

3. From LFW to Cross-Pose LFW

Our benchmark is used to achieve face verification. To simulate real world face recognition situation, we add pose difference of the same person into the dataset while keeping the identities of LFW at the same time. In this section, we first describe the process of the construction of CPLFW from collecting data to forming training and testing set in detail and then compare LFW and CPLFW.

3.1. Consruction Details

The process of building CPLFW dataset can be broken into the following steps:

1. Gathering raw images from the Internet
2. Detecting Faces.
3. Cropping and rescaling the detected faces
4. Eliminating duplicate picture
5. Judging whether labels are correct
6. Estimating the pose of each image and forming pairs of training and testing sets. Randomly selecting positive pairs and using people with the same gender and race to form negative pairs.

Gathering Images.

In order to collect images from a large number of people, Google is utilized to search face images using the identities in LFW dataset and images in the standard LFW are used as reference images to avoid finding the wrong person. 150 volunteers who are Chinese students of 18-22 years old have taken part in the collecting mission and they are asked to find two images of each person with pose difference as large as possible.

Detecting Faces.

The next step is detecting face, considering that common detection tools do not work well in large poses, we manually detect the faces. Then the image is cropped and rescaled (as described below) and saved as a separate JPEG file.

Cropping and rescaling.

For those images placed in CPLFW dataset, we use the following procedure to create them. The region obtained by human for the given face is expanded by 2 according to the maximum value of length and width. If the expanded region falls outside the original region of the image, a new image of the size equal to the size we want will be created by using black pixels to fill in the area outside the original image. The expanded image is then resized to 250×250 using the Matlab function imresize. Finally the image is saved in JPEG format.

Eliminating duplicate face photos.

Before removing the duplicate images, we need to define what is duplicates. The simplest definition is that the two images are numerically equivalent at each pixels. However, the definition ignores many situations when faces in the images are indistinguishable to the human eyes for the reason that the images collected by volunteers might have been recropped, rescaled, renormalized or variably compressed. Thus if we do not eliminate these face photos, we might

form positive pairs which are visually equivalent but differed numerically. So according to [6], we choose to define duplicates as images which are judged to have a common original source photograph. To remove duplicate images, we have the following two steps. First, a structural similarity measure [15] is used to compare all the possible couples of images from the same identity. Only the couples with a very high similarity are inspected and we delete the low quality version. To make sure that there is no duplicate image in the dataset, we then manually check pictures of each individual.

Judging whether labels are correct.

For each given subject, we pay extreme cautious to manually judge the scraped images to be truly about this celebrity or not. We use the images in the standard LFW as reference and whenever we are not sure about the label, we will use the original web page of the scraped image obtained in the gathering process and read the page content to guide the label. The rich information of the original page benefits the quality of labeling, especially for those hard cases. When the identity of the image can not be confirmed by web page, three judging people see the image together and get the final decision based on voting result. In total, we have more than 10,000 image labels which spent many hours. While we attempt to label all the pictures correctly, it is possible that certain people have been given incorrect names.

Forming training and testing sets.

In LFW view 2, it defines 10 disjoint subsets of image pairs which are suitable for cross validation. Each subset contains 300 positive pairs and 300 negative pairs. The 10 subsets are organized by their identities and each identity only appears once in certain subset. Based on it, our CPLFW dataset has been divided into 10 separate folds using the same identities contained in the LFW 10 folds. The CPLFW dataset has 2 or 3 images for each person and the name of each image is formed as follows:

$$name\_0001.jpg, name\_0002.jpg,$$

We select the positive pairs randomly. When it comes to negative pairs, to avoid attribute difference of positive pairs (same gender and race) and negative pairs (random race and gender), we first manually label the race and gender of each person in CPLFW and then select negative pairs with people who have the same gender and race randomly.

3.2. Comparison between LFW and CPLFW

In this section, we compare the standard LFW and our CPLFW. To visually view the difference, we first compare the pictures of positive pairs and negative pairs in LFW and CPLFW. After that, we compare the pose distribution between LFW and CPLFW.

The comparison of images of positive pairs and negative pairs in LFW and CPLFW are shown in Figure 2 and Figure 3. Compared to LFW, the positive pairs in CPLFW contain more obvious pose difference and the negative pairs show people of the same gender and race. This guarantees that performance on these pairs reflects a true ability to discriminate the individuals.

To compare yaw, we use Baidu Cloud Vision API to estimate the pose of images in LFW and CPLFW. The pose distribution is illustrated in Figure 4 and the pose difference distribution of positive pairs between LFW and CPLFW is shown in 5. According to the figures, the yaw distribution of images in CPLFW is more average. Also, pose difference of most positive pairs in LFW is less than 40 degrees while that of most positive pairs in CPLFW is larger. This confirms the existence of pose variation in intra-class variance of CPLFW.

If the goal of the experiment is to determine a benchmark for face verification, discrimination of the imposter pairs should be based on differences in identity only. So we select negative pairs according to race and gender attributes in CPLFW. We first label the gender and race of each person and then randomly form negative pairs using people with the same gender and race.

Also, we notice that in LFW, the image number of each person is not balanced. The database contains images of 5,749 individuals while 4,096 people have just a single image which means they can only appear in negative pairs. To increase the number of people in positive pairs so that the limited 3,000 positive pairs can better reflect the diversity of face verification in real world face recognition, each individual in CPLFW has at least 2 images.

In conclusion, there are three main differences between LFW and CPLFW. First is that pose difference has been added to intra-class variations. Second is that instead of randomly selecting negative pairs, to avoid the influence of attributes difference of positive and negative pairs, we select negative pairs using people with the same gender and race. Third is that the image number of each person in CPLFW is more balanced with 2 or 3 images for each person while the distribution of LFW is not balanced.

4. Baseline

Recently, deep convolutional neural networks trained by massive labeled outside data have reported fairly good results on face verification task of LFW benchmark. Due to the good performance, we apply two well-published convolutional neural networks to compare LFW and CPLFW, they are VGG-Face [11] and VGGFace2 [3]. The parameters of two networks are set according to the original papers and we apply the network model directly. The VGG-Face descriptors are extracted using the off-the-shelf CNN model based on the VGG-Very-Deep-16 CNN architecture as de-

Figure 2. The comparison of positive pairs in LFW and CPLFW. Compared to LFW, the positive pairs in CPLFW contain obvious pose difference.

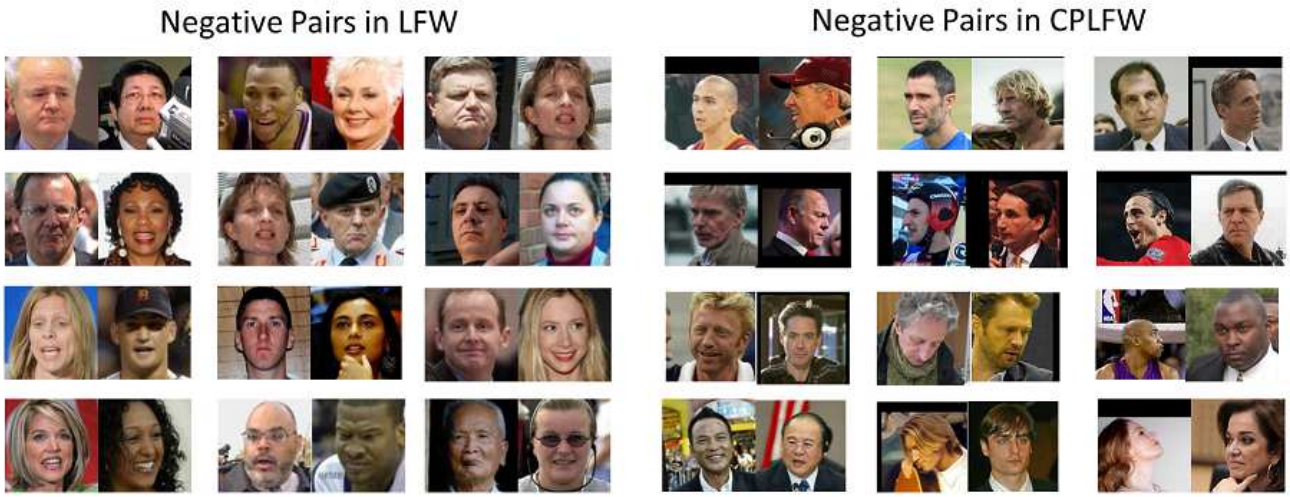

Figure 3. The comparison of negative pairs in LFW and CPLFW. Compared to LFW, the negative pairs in CPLFW show people of the same gender and race. This guarantees that performance on these pairs reflects a true ability to discriminate the individuals instead of gender or race.

scribed in [11]. The Images in the database of VGGFace2 have a wide range of pose, age, illumination and ethnicity variations of human faces. The ResNet-50 (with and without Squeeze-and-Excitation blocks [5]) Convolutional Neural Networks) are trained on VGGFace2 database, on MSCeleb-1M [4], and on their union. The network trained on VGGFace2 database leads to improved recognition performance over pose and age. Networks are learned from scratch on VGGFace2 (_scratch); Networks are first pre-trained on MS1M [4] and then fine-tuned on VGGFace2 database (_ft).

The comparison of face verification accuracy on LFW and CPLFW are reported in Table 1 and the corresponding ROC curves are shown in Figure 6.

According to the accuracy results and the ROC curves, compared to the accuracy on LFW, the accuracy drops about 12%-20% on CPLFW, which shows that by adding pose difference to intra-class variations and using negative pairs with the same gender and race, the dataset becomes difficult for face verification. Also, networks using MS1M [4] and VGGFace2 database perform better than networks using only VGGFace2 database. There are two possible rea-

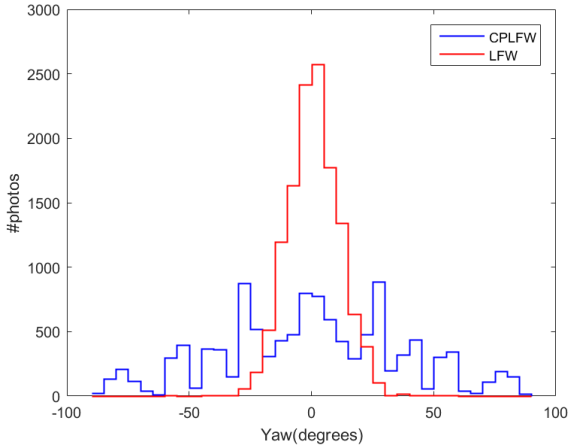

Figure 4. Compared to the images in LFW, the pose distribution of positive pairs in CPLFW is larger. This shows we successfully add pose variation to intra-class variations.

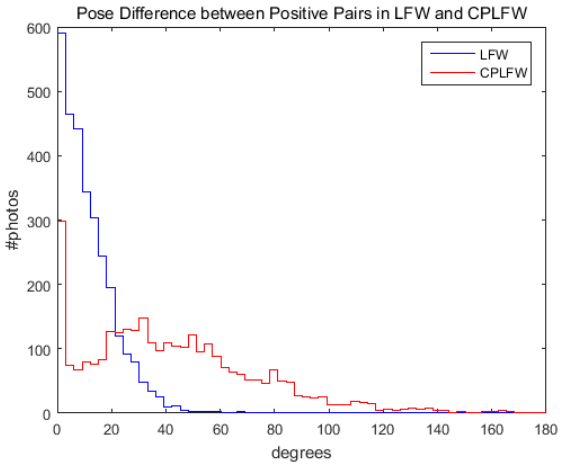

Figure 5. Compared to the positive pairs in LFW, the pose difference of positive pairs in CPLFW is larger. This shows we successfully add pose variation to intra-class variations.

Table 1. Comparison of mean verification accuracy(%) on LFW and CPLFW using deep learning approaches.

Approach	LFW	CPLFW
VGG-Face [11]	97.75%	77.90%
resnet50_scratch [3]	99.03%	79.75%
resnet50_ft [3]	99.38%	83.03%
senet50_scratch [3]	98.88%	81.07%
senet50_ft [3]	99.42%	84.45%

son for this result, one is that the two databases contain more people than one database, the other is that some people in LFW may appear in MS1M database.

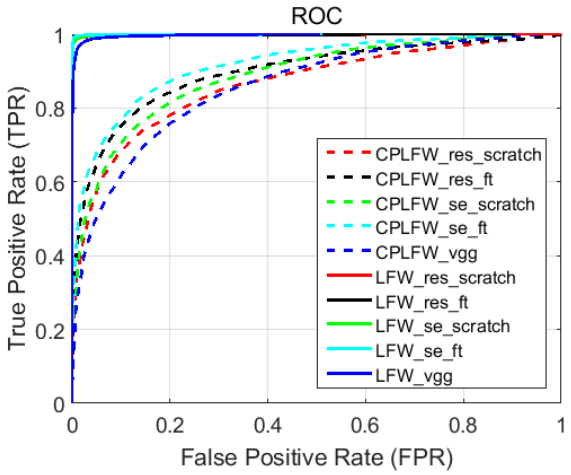

Figure 6. The ROC curves of various deep learning methods on LFW and CPLFW.

5. Discussion

In this paper, we have constructed a novel dataset according to the well-established LFW to develop face verification: the Cross-Pose Labeled Faces in-the-Wild (CPLFW) collection. The main contributions of the proposed database are: First, collecting new images according to the identity list of LFW so that each individual contains at least 2 images in the dataset. Due to the balanced distribution, more people are involved to form positive pairs to simulate the diversity of intra-class variations in real world face verification. Second, our benchmark focuses on pose difference rather than common face discrimination. Third, we concern at the attribute differences of positive pairs and negative pairs. The images of positive pairs in LFW often have same gender and same race, while the randomly selected negative pairs are often with different gender and race. To narrow the attributes difference, we randomly select people with same gender and race as negative pairs. Finally, the benchmark described in this paper provides a unified testing protocol which can easily evaluate human face verification.

We test the validity of our database by using a face detection tool and report baseline performance achieved by deep learning methods. Empirical results suggest that the CPLFW dataset provides new challenge for face verification.

References

[1] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Whos in the picture. pages 264–271, 2004.

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern*

Recognition, 2004. CVPR 2004., volume 2, pages II–848–II–854 Vol.2, June 2004.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. *MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition*, pages 87–102. Springer International Publishing, Cham, 2016.

[5] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[7] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *CoRR*, abs/1505.02108, 2015.

[9] H. W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347, Oct 2014.

[10] A. J. OToole and P. J. Phillips. Five principles for crowdsource experiments in face recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 735–741, May 2017.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, pages 41.1–41.12, 2015.

[12] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, Oct 2000.

[13] D. A. Shamma, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. J. Li. Yfcc100m: the new data in multimedia research. *Communications of the Acm*, 59(2):64–73, 2016.

[14] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 46–51, May 2002.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[16] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.