FACE RECOGNITION USING LENSLESS CAMERA

Hatef Otroshi Shahreza[1,2]*, Alexandre Veuthey*[3]*, and Sébastien Marcel*[1,4]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]ams OSRAM, Martigny, Switzerland
[4]Université de Lausanne (UNIL), Lausanne, Switzerland

ABSTRACT

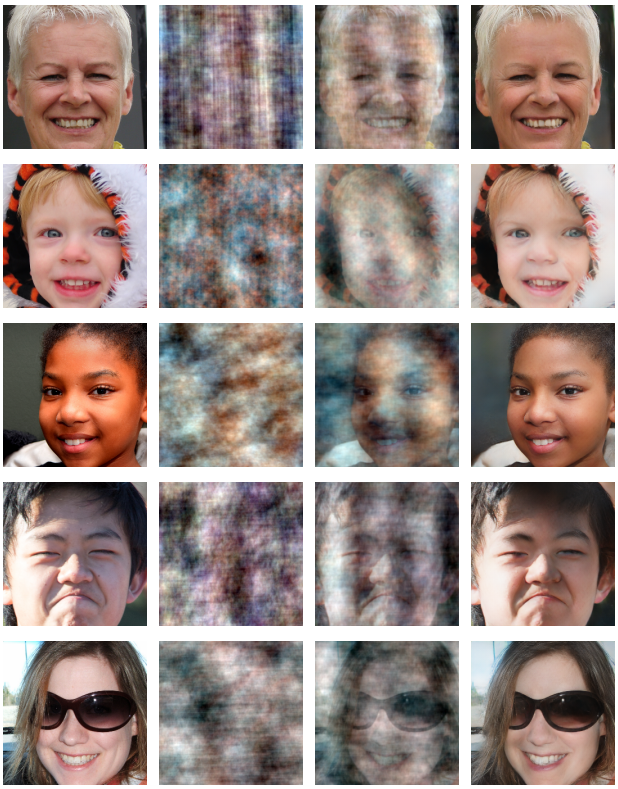
Coded aperture imaging is an emerging technique allowing thin form factor cameras that can be cheaply constructed. Many applications benefit from using such lensless cameras, such as face recognition. We propose a method for face recognition using coded aperture images that does not require retraining any component of the face recognition pipeline, but instead applies post-processing to the images with deep learning refinement so that they are compatible with existing face recognition for RGB images. We generate training data with a simulation process, based on the convolutional model of a lensless camera, and train a neural network to reconstruct face images. We train our network with a multi-term loss function to refine identity information in the reconstructed face image. We provide extensive experiments on different face recognition datasets, including LFW, CA-LFW, CP-LFW, AgeDB, FERET, and FRGC, showing the effectiveness and generalization of our proposed method. Our source code will be made available publicly to facilitate the reproducibility of our work.

***Index Terms*—** Biometrics, Coded Aperture Camera, Lensless Imaging, Face Recognition, Embeddings

1. INTRODUCTION

Originally developed for astronomical imaging, notably for X-ray and gamma-ray wavelengths [1, 2, 3, 4], coded aperture (a category of lensless) imaging is an extension of the principle of a pinhole camera. A coded aperture camera, by its construction, is much simpler to manufacture than a traditional camera with lenses or stacks of lenses. Its flat form factor can enable integration where normal cameras would not be feasible to fit. For example, integration in bidirectional (emitting and sensing) displays could be possible, thanks to flexible and transparent properties provided by a selective choice of material and design [5]. It is also possible to recover additional information about the scene, such as depth or spectral content.

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813).



(a) Original (b) Sensor (c) Deconv (d) Reconst.

Fig. 1: Samples from FFHQ and their reconstructed versions

Coded aperture cameras also show several inherent limitations. Firstly, they collect less light than lensed cameras, due to the opaque elements of the optical mask. Secondly, several physical parameters, including camera pixel pitch, mask feature size, object size, and mask-to-sensor distance, may limit the resolution of the reconstructed image. These parameters are strongly dependent on the use-cases, and may thus not be freely adjustable. Visual artifacts often appear in the reconstructed images, and contrast is related to the sparsity of the scene. Despite these limitations, new possibilities for near-infrared or visible imaging with coded aperture cameras have recently been enabled by combining their multiplexing prop-

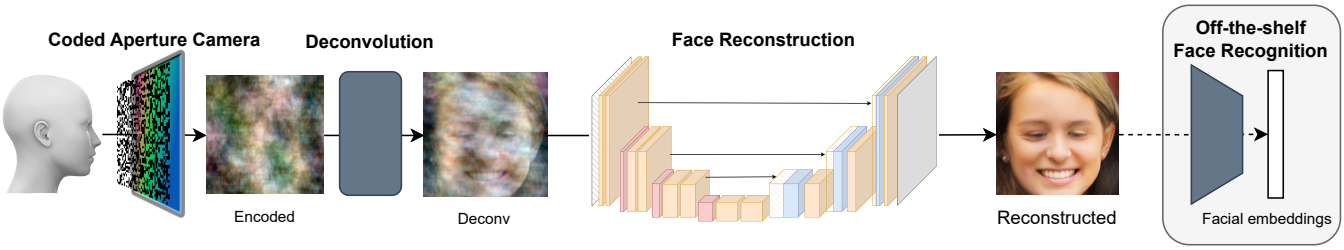


Fig. 2: Block diagram of our method: we first reconstruct the encoded sensor image with deconvolution. Then, the image reconstruction network improves the quality of the reconstructed image. The output image can be used with an off-the-shelf face recognition network.

erties with deep-learning-based models [6, 7, 8, 9, 10].

Face recognition is a challenging application to tackle, and is mostly applied on RGB images, which are taken with traditional lensed cameras. Nevertheless, [11] has shown high-accuracy face detection and verification using images taken with a FlatCam [12] device. Their approach used deep-learning-based methods to perform these tasks directly, which requires large databases for training. They compared two means of collecting data: the first is to capture a computer display showing a face image, and the second is to simulate the effect of a coded aperture mask on a clean image, purely synthetically. While they reported a better generalization for the display captured images compared to the synthetic ones, the accuracy difference is only 3%, showing the feasibility of using only synthetic data for training deep-learning-based models for face recognition. In [13], authors proposed a coded aperture camera system for face recognition that maintains high accuracy while preserving the subjects' privacy, by optimizing the mask design towards both objectives. Their approach makes privacy attacks by image reconstruction harder because of the larger open regions in the masks. In [14], high-resolution face images were generated based on extracted identity features in a coded aperture camera system.

In this paper, we propose a method for face recognition using coded aperture cameras that does not require any component of the face recognition pipeline to be retrained or fine-tuned. We achieve this by first reconstructing the encoded sensor images by deconvolution with the point spread function (PSF) of the camera. We then train a post-processing image reconstruction network, based on UNet [15], with three loss functions designed to produce a high-quality output image, in which the identity of the subject is maintained. The processed images can afterward be used in an off-the-shelf face recognition pipeline with no further adaptation. Following [11], we synthesize our sensor images with a known PSF. In contrast to [13], however, we select a mask design that maximizes the quality of the reconstructed images after deconvolution. We provide extensive experiments on different benchmarking datasets, including LFW, CA-LFW, CP-LFW, AgeDB, FERET, and FRGC, demonstrating the effectiveness of our proposed method for face recognition using lensless imaging. Fig. 1 shows sample images for lensless camera and their reconstructed face images using our proposed method.

The rest of the paper is structured as follows. Our methodology is presented in detail in Section 2. We present and discuss experiments in Section 3. Finally, the paper is concluded in Section 4.

2. METHODOLOGY

In this section, we describe our proposed method for face recognition based on lensless imaging. First, we describe how we generate synthetic sensor images in Section 2.1, and then how to reconstruct the images with deconvolution in Section 2.2. Finally, our face reconstruction network is presented in Section 2.3. Fig. 2 shows the block diagram of the proposed method.

2.1. Simulating Coded Aperture Images

To train face reconstruction networks, very large databases (starting at hundreds of thousands of images) must be collected, each original image with its associated ground-truth target. For face applications, many subjects are required for sufficient diversity and this huge scale renders any experimental data collection prohibitively expensive. Thankfully, it has been shown by [11] that generating training databases of lensless images synthetically, by applying the forward camera model on clean images, is a viable option and generalizes well to images captured with a real lensless camera.

Following the convolutional model of lensless imaging, we can obtain synthetic sensor images by convolving the face picture with the PSF of the optical system:

$$y = x * PSF, \tag{1}$$

where x is the original image, or the scene, PSF is the PSF of the camera, y is the resulting sensor image and $*$ denotes the 2D convolution. The PSF of the camera can easily be obtained by placing a point light source in front of the camera.

2.2. Image Reconstruction via Deconvolution

Using to the convolution theorem, we generate synthetic sensor images by multiplying in the Fourier domain:

$$y = \mathcal{F}^{-1}(\mathcal{F}(PSF)\mathcal{F}(x)). \quad (2)$$

To reconstruct the face image $\hat{x}$, we apply deconvolution by inverting equation 2, which would usually be expressed as a division by the PSF in the Fourier domain. However, division in Fourier has the side effect of increasing high-frequency noise. Thus, we instead multiply by the conjugate for reconstruction:

$$\hat{x} = \mathcal{F}^{-1}(\mathcal{F}(y)\overline{\mathcal{F}(PSF)}). \quad (3)$$

2.3. Face Reconstruction via Neural Network

Having a dataset of N face images (i.e., RGB images) $\{x_i\}_{i=0}^{N}$, we can generate a dataset of lensless camera and corresponding RGB images $\mathcal{D} = \{(x_i, y_i, \hat{x})_i\}_{i=0}^{N}$ using Eqs. 2,3, where y_i and $\hat{x}_i$ are image by lensless camera and reconstructed images using Eq. 3 (deconvolution), respectively. To reconstruct face images from the reconstructed images after deconvolution, we use a convolutional neural network based on UNet [15] and train it with a multi-term loss function, including:

- *Mean Squared Error (MSE):* We use the Mean Squared Error (MSE) loss term using the square of ℓ_2-norm of the reconstruction error to minimize the reconstruction error of the generated face image:

  $$\mathcal{L}_{\text{MSE}}(x, \hat{x}_{\text{UNet}}) = ||x - \hat{x}_{\text{UNet}}||_2^2, \quad (4)$$

  where $\hat{x}_{\text{UNet}}$ is the reconstructed face image by our network.

- *Learned Perceptual Image Patch Similarity (LPIPS):* To further improve reconstruction quality, we use the Learned Perceptual Image Patch Similarity (LPIPS) [16] loss:

  $$\mathcal{L}_{\text{LPIPS}}(x, \hat{x}_{\text{UNet}}) = ||P(x) - P(\hat{x}_{\text{UNet}})||_2^2, \quad (5)$$

  where $P(.)$ is a pretrained feature extractor model based on VGG-16 [17].

- *ID loss*: We also use a pretrained face recognition model $F(.)$ and minimize the ℓ_2-norm of the difference between the embeddings extracted from the original x and reconstructed $\hat{x}$ face images:

  $$\mathcal{L}_{\text{ID}}(x, \hat{x}_{\text{UNet}}) = ||F(x) - F(\hat{x}_{\text{UNet}})||_2^2, \quad (6)$$

We use a weighted summation of these loss terms as our total loss:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{LPIPS}} + \beta\mathcal{L}_{\text{ID}}, \quad (7)$$

where α and β are hyperparamters. We experimentally found that $\alpha = 1$ and $\beta = 0.05$ perform the best and we use these values for our final loss function. We train our face reconstruction network using the Adam [18] optimizer with the initial learning rate of 0.1.

3. EXPERIMENTS

3.1. Experimental Setup

Mask Design: Since our method relies on an existing face recognition pipeline, we choose a mask design that provides a quality as high as possible after deconvolution. To this end, we simply use a randomly distributed binary mask, of 256×256 features. This size was chosen to be close to traditionally used sizes for face recognition models, and also to be easily fabricated, even though all data in our experiments is synthetically generated.

Databases: We use the Flickr-Faces-HQ (FFHQ) [19] dataset for our training, which consists of 70,000 face images. We randomly split this dataset into train (90%) and validation (10%) for training our face reconstruction network. We evaluate the trained model on four different benchmarking datasets, including Labeled Faces in the Wild (LFW) [20], Cross-age LFW (CA-LFW) [21], Cross-Pose LFW (CP-LFW) [22], AgeDB-30 [23], Face Recognition Technology (FERET) [24], and Face Recognition Grand Challenge (FRGC) [25] datasets. To maintain consistency with prior works, we report recognition accuracy values on the LFW, CA-LFW, CP-LFW, and AgeDB-30 datasets, and report receiver operating characteristic (ROC) curve for evaluation on the FERET and FRGC datasets.

Implementation Details: We use ArcFace [26] with IResnet100 backbone as the RGB face recognition model in our experiments. We used the PyTorch package in our implementations. The source code of our experiments is publicly available to help reproduce our results[1].

3.2. Analysis

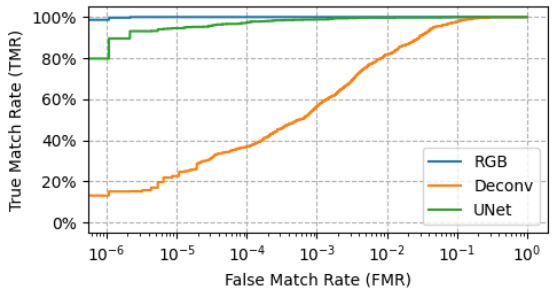
We train our face reconstruction network as described in Section 2 using the FFHQ dataset. Then, we evaluate the performance of our method on LFW, CA-LFW, CP-LFW, and AgeDB-30 datasets. Table 1 compares the recognition performance when applying ArcFace as an off-the-shelf face recognition model on deconvolution of images from the coded-aperture camera and our face reconstruction network. As the results in this table show our proposed face reconstruction network enhances identity information in reconstructed face images. In addition, the results in this table show the effectiveness of our method and generalization in a cross-dataset evaluation.

Benchmarking our method on the FERET [24] and FRGC [25] databases, shown as a receiver operating characteristic (ROC) curve in Fig. 3, also shows significant improvements of True Match Rate (TMR) across the range of False Match Rates (FMR) compared to the bare deconvolution images.

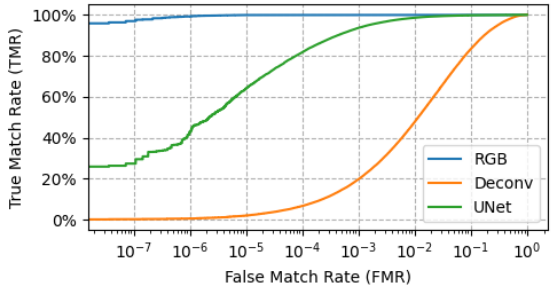
[1] Available at https://gitlab.idiap.ch/biometric/code.face_rec_lensless

Table 1: Face recognition accuracy for different types of images.

Image	Type	LFW	CA-LFW	CP-LFW	AgeDB
Original	RGB	99.85%	95.63%	91.80%	98.20%
Reconst.	Deconv	81.55%	70.28%	61.83%	71.90%
	UNet	**94.22%**	**86.95%**	**75.52%**	**87.57%**



(a) FERET



(b) FRGC

Fig. 3: Receiver operating characteristic (ROC) curves on FERET and FRGC databases

3.3. Ablation Study

Loss Function: To investigate the effect of each loss term in our proposed method we apply an ablation study and evaluate the performance of our method using different loss functions. As the results in Table 2 show each loss term in training our network improves the recognition accuracy of our method. In particular, ID loss improves the recognition accuracy of our model. The LPIPS term also contributes to the visual quality of reconstructed face images.

Mask Design: To study the effect of parameters in the mask design, we implement an ablation study and evaluate the performance of our method with different mask parameters.

Table 2: Ablation study on the effect of loss function

Loss	LFW	CA-LFW	CP-LFW
$\mathcal{L}_{\text{MSE}}$	92.25%	83.15%	72.85%
$\mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{LPIPS}}$	91.63%	83.28%	72.03%
$\mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{LPIPS}} + \beta\mathcal{L}_{\text{ID}}$	**94.22%**	**86.95%**	**75.52%**

Table 3: Ablation study on the effect of number of mask features

No.	LFW	CA-LFW	CP-LFW
32	94.37%	84.05%	73.20%
64	**95.15%**	**88.58%**	**76.75%**
128	94.03%	86.82%	75.53%
256	94.22%	86.95%	75.52%

Table 4: Ablation study on the effect of mask sparsity

Sparsity	LFW	CA-LFW	CP-LFW
30%	**96.18%**	**88.33%**	**76.47%**
40%	94.32%	87.02%	75.90%
50%	94.22%	86.95%	75.52%
60%	93.78%	86.32%	75.15%
70%	94.42%	86.60%	75.78%

First, we explore the effect of number of features in sensor mask for 32, 64, 128, and 256 features, shown in Table 3. As the results in this table show 64 features is the best overall performing choice, although the difference is not very large, and there is a sharp decrease in performance with 32 features, indicating that too few features is not useful. Using a random distribution for the mask pattern allows us to freely select its sparsity as well, contrarily to a MURA pattern, which is fixed at 50%. Therefore, as a new experiment, we explore 30%, 40%, 50%, 60%, and 70%. The results are reported in Table 4. While the best performing parameter is 30%, a low sparsity mask will in practice let less light through to the sensor, which is less desirable in practice. We should note that there are still some paramters which may affect the performance in real applications, such as mask-to-sensor distance, pixel/feature size, or illumination, which can be studied in future work.

4. CONCLUSION

In this paper, we proposed a new method for face recognition using coded aperture cameras. In our proposed method, we reconstruct RGB face images from coded aperture images and use an off-the-shelf face recognition model. Therefore, our proposed method does not require any component of the face recognition pipeline to be retrained or fine-tuned. We first reconstruct sensor images by deconvolution with the point spread function (PSF) of the camera. Then, we train a post-processing image reconstruction network, based on UNet, using a multi-term loss function. We synthesized our training dataset to train our model using a known PSF, which is shown that can be used in real-world cameras and can be manufactured. We provide extensive experiments on different benchmarking datasets, demonstrating the effectiveness of our proposed method for face recognition using lensless imaging and generalization on different datasets.

5. REFERENCES

[1] RH Dicke, "Scatter-hole cameras for x-rays and gamma rays," *Astrophysical Journal, vol. 153, p. L101*, vol. 153, pp. L101, 1968.

[2] E. E. Fenimore and T. M. Cannon, "Coded aperture imaging with uniformly redundant arrays," *Appl. Opt.*, vol. 17, no. 3, pp. 337–347, Feb 1978.

[3] Stephen R. Gottesman and E. E. Fenimore, "New family of binary arrays for coded aperture imaging," *Appl. Opt.*, vol. 28, no. 20, pp. 4344–4352, Oct 1989.

[4] P.T Durrant, M Dallimore, I.D Jupp, and D Ramsden, "The application of pinhole and coded aperture imaging in the nuclear environment," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 422, no. 1, pp. 667–671, 1999.

[5] Oliver Bimber and Alexander Koppelhuber, "Toward a flexible, scalable, and transparent thin-film camera," *Proceedings of the IEEE*, vol. 105, no. 5, pp. 960–969, 2017.

[6] Kristina Monakhova, Joshua Yurtsever, Grace Kuo, Nick Antipa, Kyrollos Yanny, and Laura Waller, "Learned reconstructions for practical mask-based lensless imaging," *Opt. Express*, vol. 27, no. 20, pp. 28075–28090, Sep 2019.

[7] Salman Siddique Khan, Varun Sundar, Vivek Boominathan, Ashok Veeraraghavan, and Kaushik Mitra, "FlatNet: Towards photorealistic scene reconstruction from lensless measurements," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[8] Vivek Boominathan, Jesse K. Adams, Jacob T. Robinson, and Ashok Veeraraghavan, "Phlatcam: Designed phase-mask based thin lensless camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 7, pp. 1618–1629, 2020.

[9] Diptodip Deb, Zhenfei Jiao, Ruth Sims, Alex B. Chen, Michael Broxton, Misha B. Ahrens, Kaspar Podgorski, and Srinivas C. Turaga, "Fouriernets enable the design of highly non-local optical encoders for computational imaging," 2022.

[10] Oliver Kingshott, Nick Antipa, Emrah Bostan, and Kaan Akşit, "Unrolled primal-dual networks for lensless cameras," *Opt. Express*, vol. 30, no. 26, pp. 46324–46335, Dec 2022.

[11] Jasper Tan, Li Niu, Jesse K. Adams, Vivek Boominathan, Jacob T. Robinson, Richard G. Baraniuk, and Ashok Veeraraghavan, "Face detection and verification using lensless cameras," *IEEE Transactions on Computational Imaging*, vol. 5, no. 2, pp. 180–194, 2019.

[12] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk, "Flatcam: Thin, lensless cameras using coded aperture and computation," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 384–397, 2016.

[13] Thuong Nguyen Canh, Trung Thanh Ngo, and Hajime Nagahara, "Human-imperceptible identification with learnable lensless imaging," 2023.

[14] Hatef Otroshi Shahreza, Alexandre Veuthey, and Sébastien Marcel, "Towards high-resolution face image generation from coded aperture camera," *IEEE Sensors Letters*, 2023.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[17] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[18] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California., USA, May 2015.

[19] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," *arXiv preprint arXiv:1812.04948*, 2018.

[20] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[21] Tianyue Zheng, Weihong Deng, and Jiani Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[22] Tianyue Zheng and Weihong Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[23] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

[24] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[25] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, June 2005, vol. 1, pp. 947–954 vol. 1.

[26] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.