Toward High-Resolution Face Image Generation From Coded Aperture Camera

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

[1] *Biometrics Security and Privacy Group, Idiap Research Institute, Martigny, Switzerland*
[2] *School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[3] *Innovation Office, ams OSRAM, Martigny, Switzerland*
[4] *School of Criminal Justice, Université de Lausanne (UNIL), Lausanne, Switzerland*
[*] *Graduate Student Member, IEEE*
[**] *Senior Member, IEEE*

Manuscript received June 30, 2023; revised 7 September 2023; accepted 8 September 2023.

Abstract—Coded aperture cameras can be manufactured cheaply, have a very thin form-factor, and may be transparent and flexible; thus providing easy-to-integrate and compact image sensors. However, limitations in reconstructed image quality and resolution have impeded the growth of their applications. We propose a method to generate high-resolution face images from low-resolution coded aperture sensor snapshots. Using the point spread function of the coded aperture camera, we generate a set of training images to train an image enhancement network. We then apply a face recognition model to extract facial templates and project them into the intermediate latent space of a face generator network to generate high-resolution (i.e., 1024×1024) face images. Our experimental results show significant retention of the subject's identity in the generated high-resolution face images. Our cross-dataset evaluation shows the generalization of our method on other datasets for generating high-resolution face images. To our knowledge, this is the first paper for generating high-resolution face images from coded-aperture imaging. The source code of our experiments is publicly available to facilitate the reproducibility of our work.

Index Terms—coded aperture camera, deep neural networks, face generation, high-resolution, lensless imaging, machine vision

I. INTRODUCTION

Coded aperture imaging was initially developed for astronomical imaging in X-ray and gamma-ray wavelengths [1]–[5], where optical elements, such as lenses and mirrors, are impossible or prohibitively expensive to manufacture. By their construction, coded aperture (lensless) cameras have several advantages over traditional cameras. Coded aperture masks are far simpler to manufacture than lenses or stacks of lenses, and the flatness of the assembled device results in a more compact package. Furthermore, flexibility and transparency can even be obtained with a selective choice of material [6], allowing integration in bidirectional (sensing and emitting) displays. In addition to visible imaging, the propagation of light through the mask and resulting multiplexing offers the possibility of encoding additional scene content, such as depth or spectral information.

On the other hand, coded aperture imaging also suffers from some limitations. The light collection is reduced by the blocking elements of the masks. In addition, the reconstructed images have low resolution caused by several parameters in these sensors, including: mask feature size, camera pixel pitch, mask-to-sensor distance, and object size, some of which may be fixed by use-case constraints. Moreover, the reconstructed images are also subject to visual artefacts and result in low-contrast images. For these limitations, coded aperture sensors are not best seen as direct replacements for traditional cameras. However, using light multiplexing properties in coded aperture cameras along with deep-learning-based models enables new possibilities for visible or near-infrared imaging [7]–[11].

(a) Original (b) Sensor (c) Deconv (d) High-res

Fig. 1: Samples from FFHQ and their reconstructed versions

Recent works have shown face recognition as a realistic application for coded aperture cameras [12], [13]. For example, [12] performed high-accuracy face detection and verification with FlatCam [14], by

Corresponding author: H. Otroshi Shahreza (e-mail: hatef.otroshi@epfl.ch).
Associate Editor: K. Ozanyan
Digital Object Identifier 10.1109/LSENS.2023.3315248

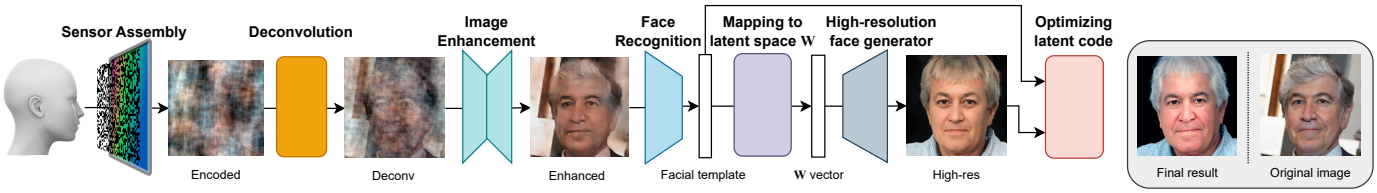

Fig. 2: Block diagram of our proposed method: we first reconstruct the image with deconvolution by the PSF of the coded aperture camera. Then, we use our image enhancement network to enhance the reconstructed image (low-resolution). Next, we extract facial templates using a face recognition model and use a mapping network to project to the intermediate latent space of a face generator network. We further optimize the latent code and generate the final high-resolution face image.

using deep-learning-based methods on coded images. They generated training images by capturing face images on displays with the lensless camera, or by applying the forward model equation to face images directly. While training on directly obtained images shows a 3% accuracy loss compared to training on display captures, they showed the feasibility of generating large-scale training databases for deep learning tasks without any hardware image acquisitions. They showed the generalization to real lensless images with a database of 88 subjects collected with a FlatCam device. In [13], a privacy-preserving face recognition system is proposed based on lensless cameras without reconstruction, where the coding effect of the camera is considered as a privacy-protection mechanism which makes template inversion difficult for an attacker.

In this paper, we propose a method to generate high-resolution (i.e., 1024×1024) face images of subjects captured with a coded aperture camera, where the output images retain the identity of the subjects. We first reconstruct the camera image by applying deconvolution with the point spread function (PSF) of the camera, and then train an image enhancement network to enhance the reconstructed image. We then extract facial templates using a face recognition model and project them to the intermediate latent space of a face generator network. We further optimize the latent codes through iterative optimization. Finally, using the face generator network, we can generate high-resolution 1024×1024 face images, where the identity information of the subjects is preserved in the high-resolution reconstructed face images. In our experiments, we synthesize sensor images with a known PSF. Fig. 1 shows sample face images from the FFHQ [15] dataset and their high-resolution reconstruction using our method. To our knowledge, this work is the first work for high-resolution face generation from coded aperture images.

The remainder of the paper is organized as follows. We explain our methodology in detail in Section II. Experiments are also presented and discussed in Section III. Finally. the paper is concluded in Section IV.

II. METHODOLOGY

In Section II-A, we describe the simulation of the coded aperture camera, to synthesize a sensor measurement y from the Point Spread Function (PSF) of the camera. To generate high-resolution face images from the sensor measurement y, we first generate an initial reconstruction $\hat{x}$ by applying deconvolution of PSF and sensor measurement as described in Section II-B. Then, as described in Section II-C, we train an image enhancement network based on the UNet structure to improve the initial reconstruction $\hat{x}$ in the same resolution, $\hat{x}_{\text{UNet}}$. Next, as described in Section II-D, we use a pre-trained face recognition model to extract facial templates t_{UNet} from $\hat{x}_{\text{UNet}}$ and then map the facial templates t_{UNet} to the intermediate latent space $\mathcal{W}$ of StyleGAN [16]. In Section II-E, we apply an optimization to further improve the mapping in the intermediate latent space $\mathcal{W}$ of StyleGAN, and finally generate the high-resolution face image through the remaining network of StyleGAN. Fig 2 depicts the block diagram of the proposed method.

A. Simulating Coded Aperture Images

Large experimental databases (hundreds of thousands of images) are costly to collect from scratch and the workaround of capturing a display is very time-consuming. As shown by [12], synthetically generating training databases for training deep-learning-based models by applying the forward camera model on clean images can be a viable option. In this case, the forward model is a simple 2D convolution of the face picture with the Point Spread Function (PSF) of the optical system:

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

where y is the sensor measurement, PSF is the PSF of the camera and x is the scene, and $*$ denotes a 2D convolution. The PSF of a camera can be obtained by measuring the impulse response of the optical system by placing a point-like source at the plane of interest.

B. Image Reconstruction via Deconvolution

Thanks to the convolution theorem, we can generate encoded face sensor images by multiplying in the Fourier domain:

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

To obtain the reconstructed face images $\hat{x}$, we apply deconvolution by inverting (2), which in theory corresponds to a division by the PSF in the Fourier domain. In practice, however, division in Fourier significantly amplifies the high-frequency noise. Therefore, we replace division with multiplication by the conjugate to have a reconstructed image:

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

C. Image Enhancement Network

Let $\{x_i\}_{i=0}^{N}$ denote a dataset of N face images (i.e., RGB images). We generate a dataset of sensor measurement $\mathcal{D} = \{(x_i, y_i, \hat{x})_i\}_{i=0}^{N}$ using Eqs. 2,3, where y_i and $\hat{x}_i$ are sensor measurement and reconstructed images using deconvolution in Eq. 3, respectively. We use this dataset to train our image enhancement network, based on the UNet [17] structure. Our image enhancement network $\mathcal{U}$ takes the reconstructed

images using deconvolution $\hat{x}$ as input and generates an enhanced image $\hat{x}_{\text{UNet}} = \mathcal{U}(\hat{x})$. We train our image enhancement network by minimizing the reconstruction error by the network using the mean squared error loss function as follows:

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

We train our image enhancement network using the Adam [18] optimizer with the initial learning rate of 0.1.

D. Mapping to the Latent Space of StyleGAN

After we trained our image enhancement network $\mathcal{U}$, we can use it to generate an enhanced image $\hat{x}_{\text{UNet}} = \mathcal{U}(\hat{x})$. Then, we use a face recognition network $\mathcal{T}$ to extract facial templates $t = \mathcal{T}(\hat{x}_{\text{UNet}})$ from enhanced image $\hat{x}_{\text{UNet}}$. Next, we use our pretrained mapping network $\mathcal{M}_{\text{t2w}}$ proposed in [19] to map facial templates to the intermediate latent space $\mathcal{W}$ of StyleGAN [16]. We can use the mapped latent code $w = \mathcal{M}_{\text{t2w}}(t)$ as an input to the synthetic network $\mathcal{S}_{\text{StyleGAN}}$ of StyleGAN and generate high-resolution face image $\hat{x}_{\text{high-res}} = \mathcal{S}_{\text{StyleGAN}}(w)$. Moreover, we can further optimize the intermediate latent code w as described in Section II-E.

E. Optimizing Latent Code

After mapping facial templates to the intermediate latent space $\mathcal{W}$ of StyleGAN, we can further optimize the intermediate latent code in the extended space $\mathcal{W}^+$ for generating the high-resolution face image by solving the following optimization:

$$w^+ = \text{argmin}_w ||\mathcal{T}(\mathcal{S}_{\text{StyleGAN}}(w)) - \mathcal{T}(\hat{x}_{\text{UNet}})||_1 \tag{5}$$

We solve this optimization with an iterative gradient-descend-based algorithm using the Adam [18] optimizer with the learning rate of 1×10^{-2} and for 20 iterations, which experimentally yield the best performance. After solving this optimization, we can generate an optimized high-resolution face image $\hat{x}_{\text{high-res,+}} = \mathcal{S}_{\text{StyleGAN}}(w^+)$. Algorithm 1 summarises our proposed method for high-resolution face generation in the inference stage.

Algorithm 1 High-resolution face image generation (inference)

Inputs:
  y : sensor measurement
  n_{itr} : number of iteration, λ : learning rate
Output:
  $\hat{x}_{\text{high-res,+}}$: High-resolution reconstructed face image
Procedure:
  Step 1: Image Reconstruction via Deconvolution
    $\hat{x} = \mathcal{F}^{-1}(\mathcal{F}(y)\overline{\mathcal{F}(PSF)})$
  Step 2: Applying Image Enhancement Network
    $\hat{x}_{\text{UNet}} = \mathcal{U}(\hat{x})$
  Step 3: Extracting Facial Templates
    $t = \mathcal{T}(\hat{x}_{\text{UNet}})$
  Step 4: Mapping to Latent Space of StyleGAN
    $w = \mathcal{M}_{\text{t2w}}(t)$
  Step 5: Optimizing Latent Code
    Set initial value of w^+ with w
    for itr **in** $\{1, \ldots, n_{\text{itr}}\}$ **do**
      cost $= ||\mathcal{T}(\mathcal{S}_{\text{StyleGAN}}(w^+)) - \mathcal{T}(\hat{x}_{\text{UNet}})||_1$
      $w^+ \leftarrow w^+ - \text{Adam}(\nabla\text{cost}, \lambda)$
    end for
  Step 6: Generating High-resolution Face Image
    $\hat{x}_{\text{high-res,+}} = \mathcal{S}_{\text{StyleGAN}}(w^+)$
End Procedure

III. EXPERIMENTS

To synthesize sensor images in our experiments, we used a purely random pattern design and opted for 256×256 image resolution. We should note that mask alignment to sensor pixels is a sensitive process when using real coded aperture cameras. In our synthetic approach, we design the mask to have 256×256 features, such that one feature matches one pixel on the sensor. Furthermore, we chose 50% sparsity in the mask features, which represents a good compromise between light input and invertibility.

In our experiments, we use the Flickr-Faces-HQ Dataset (FFHQ) dataset [15] as our dataset of RGB face images $\{x_i\}_{i=0}^N$, and generate our dataset of sensor measurement $\mathcal{D} = \{(x_i, y_i, \hat{x})_i\}_{i=0}^N$ using Eqs. 2, as described in Section II-C. The FFHQ dataset consists of 70,000 face images with variations in terms of age, ethnicity, etc. We randomly split this dataset into training and validation sets and train our image enhancement network. We use the same training set to train the mapping from the ArcFace [20] face recognition model to the intermediate latent space of StyleGAN [16] based on [19]. Fig 1 shows sample face images from the validation set of the FFHQ dataset and their corresponding high-resolution (i.e., 1024×1024) reconstructed face images with our proposed method.

(a) Original (b) Sensor (c) Deconv (d) High-res

Fig. 3: Samples from FERET and their reconstructed versions.

In addition to the FFHQ dataset and as a cross-dataset evaluation, we used FERET [21] dataset and similarly generated coded aperture camera images for the frontal images in this dataset. Then, we reconstructed high-resolution (i.e., 1024×1024) face images using our proposed method (and trained models on FFHQ). Fig.3 illustrates sample face images from the FERET dataset and their corresponding high-resolution reconstructed face images with our proposed method. In addition, Fig. 4 shows the histogram of cosine similarity of ArcFace templates of deconvolution and the reconstructed face images using our method as well as the ArcFace templates of original RGB images. This histogram also shows the cosine similarity of mated

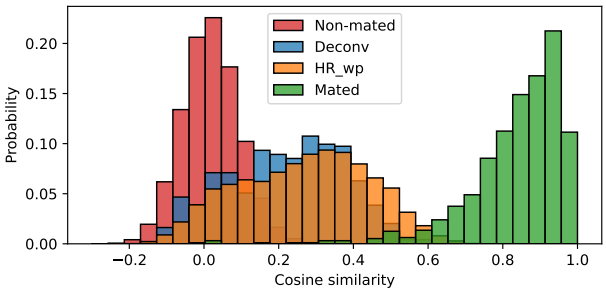

Fig. 4: Histogram of face recognition similarity scores on FERET

Table 1: Ablation Study.

Image	Reconstruction Resolution	Cosine Similarity FFHQ	Cosine Similarity FERET
Mated RGB Images (baseline)	-	N/A	0.84
Deconv ($\hat{x}$)	256×256	0.30	0.17
Enhanced ($\hat{x}_{\text{UNet}}$)	256×256	0.56	0.34
High Res ($\hat{x}_{\text{high-res}}$)	1024×1024	0.35	0.17
Optimized High Res ($\hat{x}_{\text{high-res,+}}$)	1024×1024	0.51	0.27

(different images of same subject) and non-mated (images of different subjects) pairs in this dataset. As the results in this histogram show our method improves the identity information of face images compared to deconvolution while generating high-resolution face images.

To further investigate the effect of each part in our proposed method, as an ablation study we calculate the average cosine similarity of face recognition templates extracted from images in each step in our method with the templates of original images. As the results in Table 1 show, the image enhancement network improves the identity information in the enhanced images. However, the resolution of the images is low and the images still have some artifacts (i.e., not realistic). When we generate high-resolution images $\hat{x}_{\text{high-res}}$, the identity information is reduced, but in our final optimized high-resolution images $\hat{x}_{\text{high-res,+}}$ we observe that we have comparable cosine similarity with realistic and high-resolution reconstructed images with our enhanced images. Compared to reconstruction via deconvolution, our method generates high-resolution images with more identity information. The same trend holds for our reconstruction of FFHQ (validation) and FERET datasets, which shows the generalization of our proposed method.

We should note that the source code of our experiments is publicly available[1] to facilitate the reproducibility of our work.

IV. CONCLUSION

We proposed a method to generate high-resolution (i.e., 1024×1024) face images from raw sensor images from a coded aperture camera. While direct reconstruction quality is far from traditional lensed cameras, enough information is retained in the reconstruction of sensor images, and that information can be used to generate high-resolution face images of the same subject. In this paper, we used the PSF of the coded aperture camera to generate sensor images. Then, to reconstruct images, we first used deconvolution of PSF, and enhanced our reconstruction with a neural network. Finally, we extracted facial templates from enhanced images by our image enhancement network and mapped them into the intermediate space of a face generator network. We further optimized the mapped latent codes and used them to generate high-resolution face images. To our knowledge, this is the first work on the reconstruction of high-resolution face images from coded aperture camera.

[1]https://gitlab.idiap.ch/bob/bob.paper.sensl2023_hires_codedaperture

ACKNOWLEDGEMENT

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813). The authors would like to thank Antoine Boniface (ams OSRAM) for his valuable comments and fruitful discussions.

REFERENCES

[1] R. Dicke, "Scatter-hole cameras for x-rays and gamma rays," *Astrophysical Journal, vol. 153, p. L101*, vol. 153, p. 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. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?URI=ao-17-3-337

[3] S. 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. [Online]. Available: https://opg.optica.org/ao/abstract.cfm?URI=ao-28-20-4344

[4] T. M. Cannon and E. E. Fenimore, "Coded Aperture Imaging: Many Holes Make Light Work," *Optical Engineering*, vol. 19, no. 3, p. 283, Jun. 1980.

[5] P. Durrant, M. Dallimore, I. 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. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168900298010146

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

[7] K. Monakhova, J. Yurtsever, G. Kuo, N. Antipa, K. Yanny, and L. Waller, "Learned reconstructions for practical mask-based lensless imaging," *Opt. Express*, vol. 27, no. 20, pp. 28 075–28 090, Sep 2019. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-27-20-28075

[8] S. S. Khan, V. Sundar, V. Boominathan, A. Veeraraghavan, and K. Mitra, "FlatNet: Towards photorealistic scene reconstruction from lensless measurements," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. [Online]. Available: https://doi.org/10.1109%2Ftpami.2020.3033882

[9] V. Boominathan, J. K. Adams, J. T. Robinson, and A. 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.

[10] D. Deb, Z. Jiao, R. Sims, A. Chen, M. Broxton, M. B. Ahrens, K. Podgorski, and S. C. Turaga, "Fourierets enable the design of highly non-local optical encoders for computational imaging," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 224–25 236, 2022.

[11] O. Kingshott, N. Antipa, E. Bostan, and K. Akşit, "Unrolled primal-dual networks for lensless cameras," *Opt. Express*, vol. 30, no. 26, pp. 46 324–46 335, Dec 2022. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-30-26-46324

[12] J. Tan, L. Niu, J. K. Adams, V. Boominathan, J. T. Robinson, R. G. Baraniuk, and A. Veeraraghavan, "Face detection and verification using lensless cameras," *IEEE Transactions on Computational Imaging*, vol. 5, no. 2, pp. 180–194, 2019.

[13] T. N. Canh, T. T. Ngo, and H. Nagahara, "Human-imperceptible identification with learnable lensless imaging," *IEEE Access*, 2023.

[14] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk, "Flatcam: Thin, lensless cameras using coded aperture and computation," *IEEE Transactions on Computational Imaging*, vol. 3, no. 3, pp. 384–397, 2016.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[16] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

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

[19] H. O. Shahreza and S. Marcel, "Face reconstruction from facial templates by learning latent space of a generator network," *Advances in Neural Information Processing Systems*, 2023.

[20] J. Deng, J. Guo, X. Niannan, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] 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*, vol. 22, no. 10, pp. 1090–1104, 2000.