Supplementary Material
Learning to Describe Scenes via Privacy-aware Designed Optical Lens

Paula Arguello, Jhon Lopez, *Student Member, IEEE*, Karen Sanchez, Carlos Hinojosa, *Member, IEEE*, Fernando Rojas-Morales, Henry Arguello, *Senior Member, IEEE*

I. Introduction

This supplementary document provides additional experiments, visualizations, and implementation details of our work. Specifically, we include the following:

1) Privacy validation
   - Face recognition
   - VISPR protocol
   - Phase recovery robustness
2) End-to-end ablation study
3) Optical diagram of the hardware implementation
4) Double-LSTM approach

II. Privacy Validation

A. Face recognition

In addition to evaluating resistance to adversarial and facial recognition attacks, presented in section 4 of the main manuscript, we performed additional experiments to evaluate the resistance of our privacy methods against a facial recognition network. For this proof, we use the AdaFace network [4] and address the performance on three datasets: the Cross-Age Labeled Faces in the Wild (CALFW) [1], the Cross-Pose Labeled Faces in the Wild (CPLFW) [2], and the Labeled Faces in the Wild (LFW) [3]. Those datasets contain face images labelled with person names. The plots in Fig. 2 show that privatizing the face images with our proposed system resulted in a decrease in average precision (AP) for face recognition by approximately 20% across all evaluated datasets, one per plot, compared to results using the original face images.

B. VISPR protocol

Furthermore, we have developed additional experiments following the VISPR protocol to address privacy concerns related to attributes in images that can be exposed as nudity, and skin color, among others.

In Fig. 1, we have included the average Precision-recall curves of six different specific attributes from this experiment: culture, age, weight, credit card, occupation, and religion. The

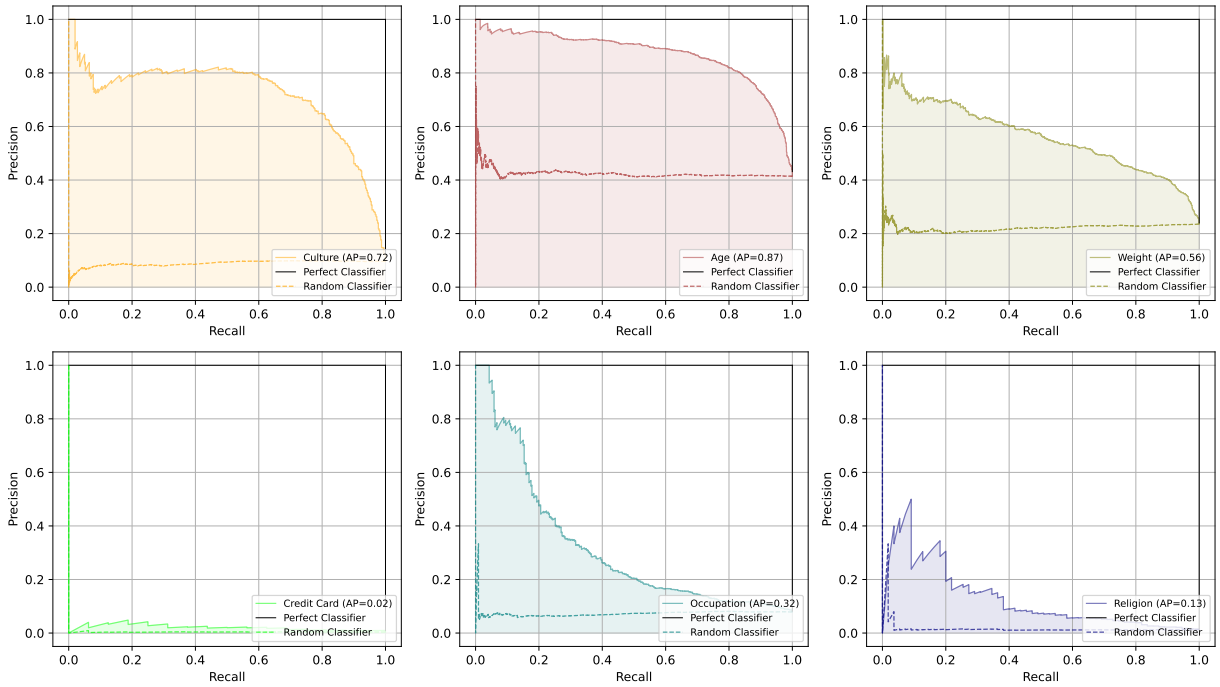

Fig. 1: Privacy attribute evaluation: Precision-recall curves of culture, age, weight, credit card, occupation, and religion attribute recognition on the VISPR dataset private through our proposed method.

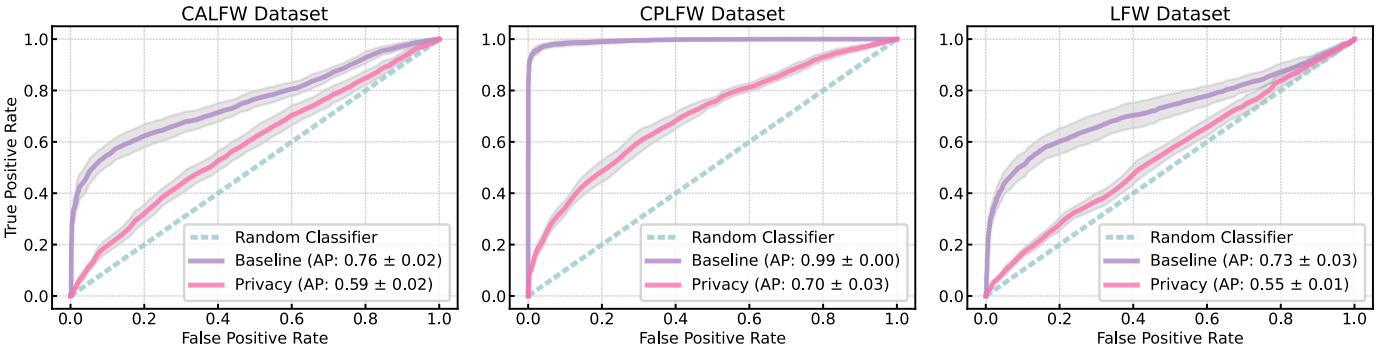

Fig. 2: ROC curves for a face detection model across three datasets: CALFW [1], CPLFW [2], and LFW [3]. "Baseline" represents the performance using standard RGB images, while "Privacy" depicts outcomes using private images distorted by our optimized lens. Additionally, the performance of a "Random Classifier" is shown for comparison purposes.

results of this recent experiment illustrate the effectiveness of our approach in safeguarding these attributes against potential privacy breaches.

C. Phase-Recovery robustness

Moreover, to confirm that our method remains effective after using an iterative algorithm for phase recovery, we conducted experiments with three recovery algorithms: Reweighted AmpFlow (RAF) [5], Truncated AmpFlow (TAF) [6], and Truncated Wirtinger Flow (TWF) [7]. The results, which can be seen in Fig. 3 involve recovering each color channel of the outcome phase of our distorted lens. As expected, the results show that the distorted images cannot be recovered.

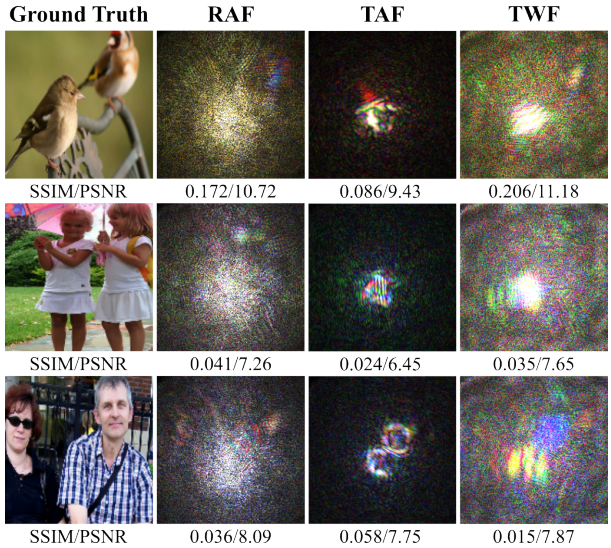

Fig. 3: Qualitative and quantitative results on phase recovery algorithms: RAF, TAF, and TWF. Below each image, the SSIM and PSNR values are provided, comparing the original images (Ground Truth) with the reconstructions.

III. End-to-end ablation studies

To further demonstrate the advantages of the end-to-end training process, we show some experiments and results when using a modular approach consisting of two stages: (1) optical encoder optimization and (2) image captioning network optimization. Specifically, in the first stage of the modular approach, we only learn the Zernike coefficients of our privacy-preserving lens without considering the image captioning network and maximizing the mean square error between the original image and the captured sensor image. To avoid significant degradation of the lens that leads to information loss and poor feature representation, we stop the optimization of the optical encoder when the point spread function (PSF) of the lens resembles the PSF obtained with the end-to-end training approach. Once the optimization of the first stage is finished, we train the image captioning network using as input the privacy-preserving images provided by the already optimized optical encoder. We present qualitative and quantitative results with both approaches, the end-to-end (ours) and the modular (two-stage), in Fig. 4 and Table I, respectively. As observed from the results, learning the privacy-preserving lens and image captioning parameters end-to-end significantly improves the performance leading to an optical encoder which better preserves privacy and encodes useful information from the original scene for the image captioning network (decoder).

IV. Optical diagram of the hardware implementation

An optical diagram corresponding to the hardware setup from Fig. 6 of the main manuscript is presented in this section. This setup, depicted in Fig. 6, incorporates a 100 mm objective lens to direct the input light rays, which are encoded by a digital micromirror device (DM, Thorlabs DMP40-P01). This device emulates the refractive lens designed by our end-to-end method. A Fourier transforming lens (Thorlabs AC254-075-A-ML), referred to as L1 in Fig 6, is placed at a distance of $1f = 100\ mm$ from the DMD. A beam splitter (BS, Thorlabs CCM1-BS013) then redirects the wavefront-encoded light from the DM to another Fourier transforming lens, L2, which subsequently directs the light to the sensor for acquisition.

V. Double-LSTM approach

In this experiment, we validate our work with another captioning network. Specifically, we have implemented a dual

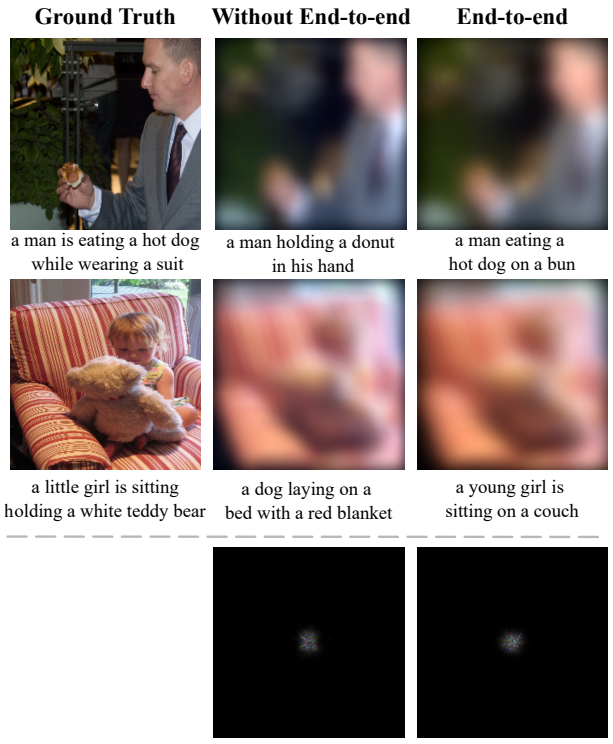

Fig. 4: Qualitative comparison of End-to-end vs. two-stage (without end-to-end) optimization approach. From left to right: Ground Truth, sensor image from the optical encoder optimized without end-to-end, and optical encoder trained end-to-end (ours). Below, the corresponding Point Spread Functions (PSF) are displayed for each approach.

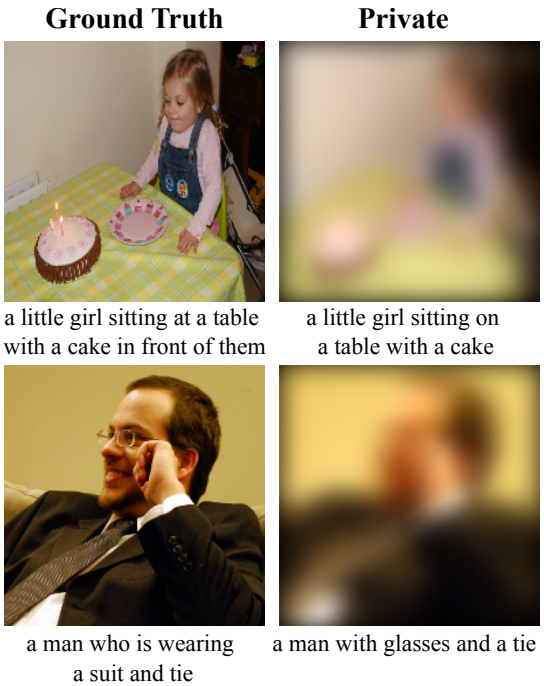

Fig. 5: Qualitative results when using double LSTM as image captioning network (decoder). The "Private" column shows the image acquired with our optimized lens and the corresponding image captioning result.

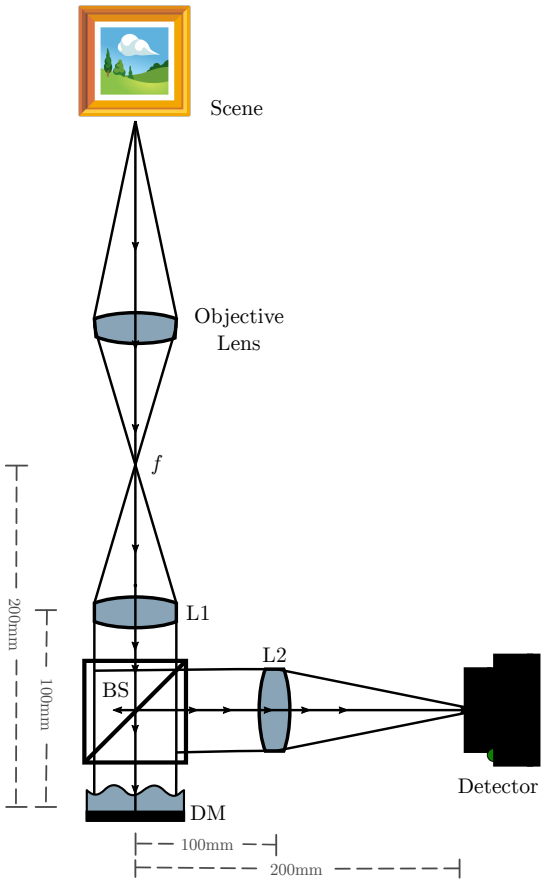

Fig. 6: Optical diagram of our experimental hardware setup. In the figure, DM stands for deformable mirror and BS for beam splitter.

LSTM architecture [8] as the decoder. In Table II, we present results using original images (✗ Private) and images processed through our optimized lens (✓ Private). These results include the COCO and Flickr8k datasets for generating captions with the dual LSTM.

The table presents image captioning metrics to demonstrate the effectiveness of our approach across various models. Notably, the results using the dual LSTM architecture [8] surpass those in the main manuscript in terms of captioning accuracy. Additionally, Fig. 5 illustrates qualitative results, showing a comparison between the original images with their ground truth captions and the privacy-enhanced images with their predicted captions.

REFERENCES

[1] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[2] T. Zheng and W. 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, p. 5, 2018.

[3] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2879–2886.

TABLE I: End-to-end (✔ E2E) vs. two-stage (✗ E2E) approach comparison on the COCO validation set.

Dataset	E2E	Bleu-1(%)↑	Bleu-2	Bleu-3	Bleu-4	METEOR↑	Rouge↑	Cider↑
COCO	✗	63.7	45.5	31.5	22.0	25.6	36.4	84.1
	✔	**68.9**	**51.3**	**37.3**	**27.0**	**28.1**	**38.1**	**88.5**

TABLE II: Quantitative results when using double LSTM as the image captioning network (decoder) in our approach evaluated on the COCO and Flickr8k datasets.

Dataset	Private	Bleu-1(%)↑	Bleu-2	Bleu-3	Bleu-4	METEOR↑	Rouge↑	Cider↑
COCO	✗	73.2	56.7	43.4	33.3	29.0	39.4	101.2
	✔	68.9	51.7	38.5	29.0	26.7	37.6	89.0
Flickr8k	✗	65.5	48.9	35.5	25.0	26.3	35.9	65.4
	✔	64.6	45.8	32.0	21.6	26.2	35.5	59.3

[4] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.

[5] G. Wang, G. B. Giannakis, Y. Saad, and J. Chen, "Phase retrieval via reweighted amplitude flow," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2818–2833, 2018.

[6] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 773–794, 2017.

[7] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.