Poster: Natural Face De-Identification

Tao Li
Dept. of Computer Science, Purdue University
taoli@purdue.edu

Abstract

In an effort to enhance privacy while balance usability and reality of facial images, we propose a novel framework called AnonymousNet, which encompasses four stages: (i) facial attribute prediction by leveraging a deep Convolutional Neural Network (CNN); (ii) facial semantic obfuscation directed by privacy metrics; (iii) natural image synthesis using a Generative Adversarial Network (GAN); and (iv) universal adversarial perturbation against malicious detectors. Not only do we achieve the state-of-the-arts in terms of image reality and attribute prediction accuracy, we are also the first to show that facial privacy is measurable, can be factorized, and accordingly be manipulated in a photo-realistic fashion to adapt to different privacy requirements and application scenarios. Experiments further demonstrate the effectiveness of the proposed framework.

1 Introduction

With billions of personal images being generated from social media and cameras of all sorts on a daily basis, security and privacy are unprecedentedly challenged. Although extensive attempts have been made, existing face image de-identification techniques are either insufficient in photo-reality or incapable of balancing privacy and usability qualitatively and quantitatively, i.e., they fail to answer counterfactual questions such as "is it private now?", "how private is it?", and "can it be more private?" In this paper, we propose the AnonymousNet, with an effort to systematically address these long-standing issues and preserve privacy in a natural, measurable, and controllable manner. We leave more discussions and results in the full paper [2].

(a) Original **(b)** Blurring **(c)** Pixelation **(d)** Ours

Figure 1: Comparison of face obfuscation methods.

2 The AnonymousNet

Stage-I: Facial Attribute Prediction Using Convolutional Neural Networks

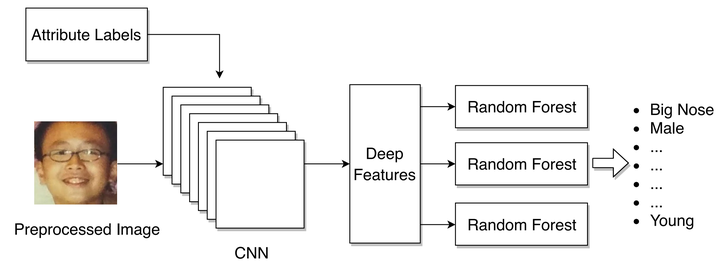

Figure 2: Facial attribute prediction pipeline.

Stage-II: Privacy-Aware Facial Semantic Obfuscation

Facial attributes are selected subject to t-closeness, i.e., the distribution S_E of any attribute E is close to its distribution S in the entire dataset. We further introduce a stochastic perturbation towards ϵ-differential privacy.

Stage-III: Natural De-Identified Face Generation Using GAN

After obtaining facial attributes that satisfies privacy constraints computed from previous steps, we train a Generative Adversarial Network (GAN) for face attributes translation, which is designed as two players, D and G, playing a minmax game with adversarial loss:

$$L_{adv} = \mathbb{E}[\log(D(\mathbf{x}))] + \mathbb{E}[\log(1 - D(G(\mathbf{x})))] \quad (1)$$

where generator G is trained to fool discriminator D, who tries to distinguish real images from adversarial ones.

Stage-IV: Adversarial Perturbation

We further introduce a universal perturbation [3] adjusted by parameter ϵ added to synthesized images, which tricks malicious detectors while preserves perceptual integrity.

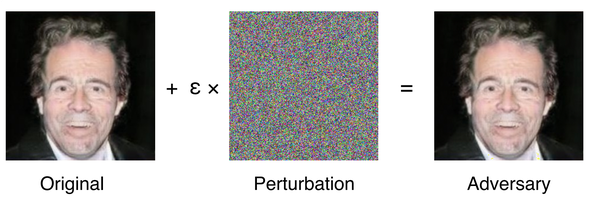

Figure 3: An example of adversarial perturbation.

3 Experiment

3.1 Dataset

We adopt the CelebA Dataset to train the facial attribute prediction model in Stage-I, which contains $202,599$ images and each image has 40 attribute labels of boolean values (e.g., `Big Nose`, `Big Lips`, and `Narrow Eyes`).

3.2 Image Preprocessing

Before feeding the data into our deep models, we perform data preprocessing for each images in the datasets. We deploy a Deep Alignment Network (DAN) [1] to obtain facial landmarks and accordingly align faces and crop images.

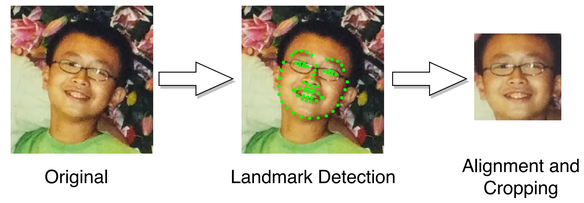

Figure 4: Image preprocessing pipeline.

3.3 Evaluation by Visual Turing Tests

Figure 5: Experimental results. In each pair, left is the original image and right is the synthesized result with an altered identity.

References

[1] M. Kowalski, J. Naruniec, and T. Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 88–97, 2017.

[2] T. Li and L. Lin. AnonymousNet: Natural face de-identification with measurable privacy. *arXiv preprint arXiv:1904.12620*, 2019.

[3] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.