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.**

I. 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. Figure 1 compares our approach with existing methods. In the rest of this poster paper, we outline our approach in Section II, and evaluate experimental results in Section III. We leave more discussions and results in the upcoming full paper [5].

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

Fig. 1: Comparison of face obfuscation methods.

II. The AnonymousNet

Stage-I: Facial Attribute Prediction

We leverage a deep Convolutional Neural Network for facial attribute prediction, using sigmoid cross-entropy loss:

$$L = -\frac{1}{n}\sum[y \ln a + (1-y)\ln(1-a)] \tag{1}$$

where y is labels and a is outputs. After extracting deep features from the fully connected layer, we train 40 random forest classifiers for attribute estimation and achieve a state-of-the-art accuracy. Figure 3 outlines the pipeline.

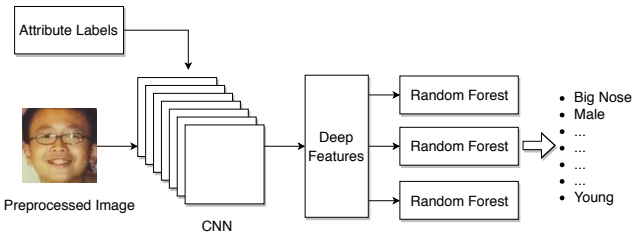

Fig. 3: Facial attribute prediction pipeline.

Stage-II: Privacy-Aware Facial Semantic Obfuscation

Facial attributes are selected subject to t-closeness [4], 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

After obtaining facial attributes that satisfies privacy constraints computed from previous steps, we train a Generative Adversarial Network (GAN) [2] for face attributes translation, which is designed as two players, D and G, playing a minmax game with adversarial loss:

$$L_{adv} = \mathbb{E}[\log(D(\mathbf{x}))] + \mathbb{E}[\log(1 - D(G(\mathbf{x})))] \tag{2}$$

where generator G is trained to fool discriminator D, who tries to distinguish real images from adversarial ones. We use Wasserstein loss [1] here to expedite the training process:

$$\mathcal{L}_{adv} = \mathbb{E}_x[D_{src}(X)] - \mathbb{E}_{x,c}[D_{src}(G(x,c))] \tag{3}$$

$$= -\lambda_{gp}\mathbb{E}_{\hat{x}}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{4}$$

where $\hat{x}$ is uniformly sampled between a pair of original and synthesized images and we set $\lambda = 10$ here.

Fig. 2: Experimental results. In each pair, left is the original image and right is the synthesized result with an altered identity.

Stage-IV: Adversarial Perturbation

As illustrated in Figure 4, we further introduce a small universal perturbation [7] adjusted by parameter ϵ added to synthesized images, which tricks malicious detectors while preserves perceptual integrity.

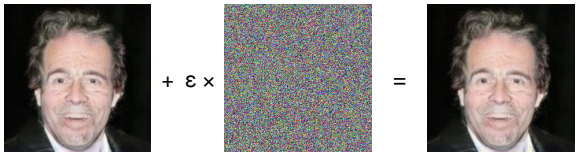

Fig. 4: An example of adversarial perturbation.

III. Experiment

A. Dataset

We adopt the Large-scale CelebFaces Attributes (CelebA) Dataset [6] to train the facial attribute prediction model in Stage-I, which contains $202,599$ images and $10,177$ identities, and each image has 40 attribute labels of boolean values (e.g., `Big Nose`, `Big Lips`, and `Narrow Eyes`).

B. 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) [3] to obtain facial landmarks, based on which we accordingly align faces and crop images. Figure 5 illustrates our approach.

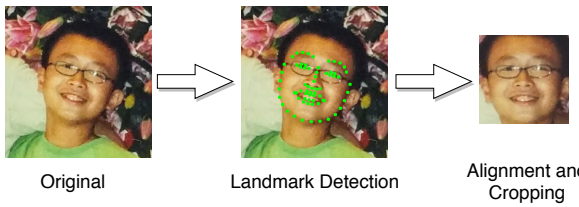

Fig. 5: Image preprocessing pipeline.

C. Evaluation

Figure 2 demonstrates experimental results, showing that identities are preserved in a perceptually natural manner; meanwhile, each pair of images still shares certain common attributes in correspondence with various privacy policies and application scenarios. Artifacts introduced by the proposed the adversarial perturbation are also shown to be visually negligible. Below are numerical metrics for reference.

	Blurring	Pixelation	Masking	Inpainting	Ours
PSNR	24.532	22.802	15.459	18.097	20.079
SSIM	0.8281	0.8226	0.8762	0.8020	0.7894
MS-SSIM	0.8842	0.8784	0.8204	0.7659	0.8650

IV. Conclusion

In this work, we propose the AnonymousNet framework to address the face de-identification problem, which is capable of balancing usability while enhancing privacy in a natural and measurable manner. The framework encompasses four stages: facial attribute estimation, privacy-metric-oriented face obfuscation, directed natural image synthesis, and adversarial perturbation. Experiments demonstrate that identities have been well obfuscated by photo-realistic alternatives of visually convincing quality, and artifacts introduced by the synthesis and the perturbation are perceptually negligible.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[3] 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.

[4] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

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

[6] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[7] 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.