A Hybrid Model for Natural Face De-Identification with Adjustable Privacy

Yunqian Wen, Bo Liu, Rong Xie, Yunhui Zhu, Jingyi Cao, and Li Song

***Abstract*—As more and more personal photos are shared and tagged in social media, security and privacy protection are becoming an unprecedentedly focus of attention. Avoiding privacy risks such as unintended verification, becomes increasingly challenging. To enable people to enjoy uploading photos without having to consider these privacy concerns, it is crucial to study techniques that allow individuals to limit the identity information leaked in visual data. In this paper, we propose a novel hybrid model consists of two stages to generate visually pleasing de-identified face images according to a single input. Meanwhile, we successfully preserve visual similarity with the original face to retain data usability. Our approach combines latest advances in GAN-based face generation with well-designed adjustable randomness. In our experiments we show visually pleasing de-identified output of our method while preserving a high similarity to the original image content. Moreover, our method adapts well to the verificator of unknown structure, which further improves the practical value in our real life.**

***Index Terms*—privacy protection, multi-level features fusion, GAN-based face generation, randomness**

I. Introduction

In such an open era, the popularity of smartphones allows people to take their life photos conveniently. Particularly, the blooming development of media and network techniques makes a vast amount of photos more approachable. At the same time, however, advanced image retrieval and face verification algorithms allow to index and recognize privacy relevant information more reliably than ever. Consequently, among those image sources exposed to the public with or without our awareness, the wide range of private information inadvertently leaked as a consequence is severely under-estimated [1]. To address this exploding privacy threat, it is crucial to study techniques that allow individuals to effectively hide the identity information and preserve the realism of the visual data.

In traditional computer vision community, privacy-enhancing technologies are mainly obfuscation-based. For example, obfuscating sensitive information like faces and numbers in an image by using approaches including blurring, Mosaic and Masking. However, researchers have shown these techniques are vulnerable [2]. Another study also shows that deep learning models can successfully identify faces in images encrypted with these techniques with high accuracies [3]. What's more, obfuscation-based approaches towards manipulating images lead to destroying the usability of images. A study shows that both blurring and blocking will impact image perception scores, and even lower scores are observed for images obfuscated by blocking [4].

New techniques and mechanisms are being applied to enhance image obfuscation. The fundamental idea is to generate a small but intentional worst-case disturbance to an original image, which misleads deep neural networks (DNN) without causing a significant difference perceptible to human eyes. The perturbed image is called an "adversarial example" and the specially generated noise is named adversarial perturbations (AP). *Liu et al.* [5] discussed the potential of AP in privacy protection in face recognition. Then *Xue et al.* [6] proposed to add adversarial perturbations to the sensitive parts of the images so the private information can be hidden while the rest parts of the images are still visible to AI. It performs well in white box attacks, but may fail in the case of unknown network models, which limits its usage.

Nowadays, the generative adversarial networks (GANs) provide an inspiring framework on generating sharp and realistic natural face image samples via adversarial training [7], therefore it has become more and more popular for new face de-identification techniques. *Wu et al.* [8] proposed a PP-GAN with specially designed verificator and regulator modules to achieve face de-identification. It is believed that GAN alone cannot distinguish the deep embedding that represents identity and that maintains the visual similarity of the rest part of the image. Consequently some auxiliary structures must be designed, which lead to a complicated network structure as well as difficult training process. Moreover, *Li et al.* [9] proposed an AnonymousNet framework, which performs de-identification in four stages. The result is remarkable, but since it uses face attributes annotations exchange to translate identity and adopts AP, the generated effect is easy to distortion and cannot deal well with black box attacks.

With the development of deep learning research, deep convolutional neural networks (CNNs) have been proven to have a strong ability to extract disentangled deep features according to demands [10]. When CNNs are trained in face recognition field, they develop a representation of the image that makes face information increasingly explicit along the processing hierarchy [11]. *Sun et al.* [12] proposed DeepID2 features which can be used to represent a face photo. *Deng et al.* [13] proposed an ArcFace to obtain highly discriminative disentangled features for face recognition. It outperforms the state-of-the-art and can be easily implemented with negligible computational overhead.

In this paper, we propose a novel two-stage approach

Yunqian Wen, Rong Xie, Yunhui Zhu, Jingyi Cao are with Shanghai Jiao Tong University. e-mail: wenyunqian, xierong, zhuyunhui, cjycaojingyi, song$_l i$@$sjtu.edu.cn$.

Bo Liu is with School of Computer Science, University of Technology Sydney, Australia. e-mail: bo.liu@uts.edu.au

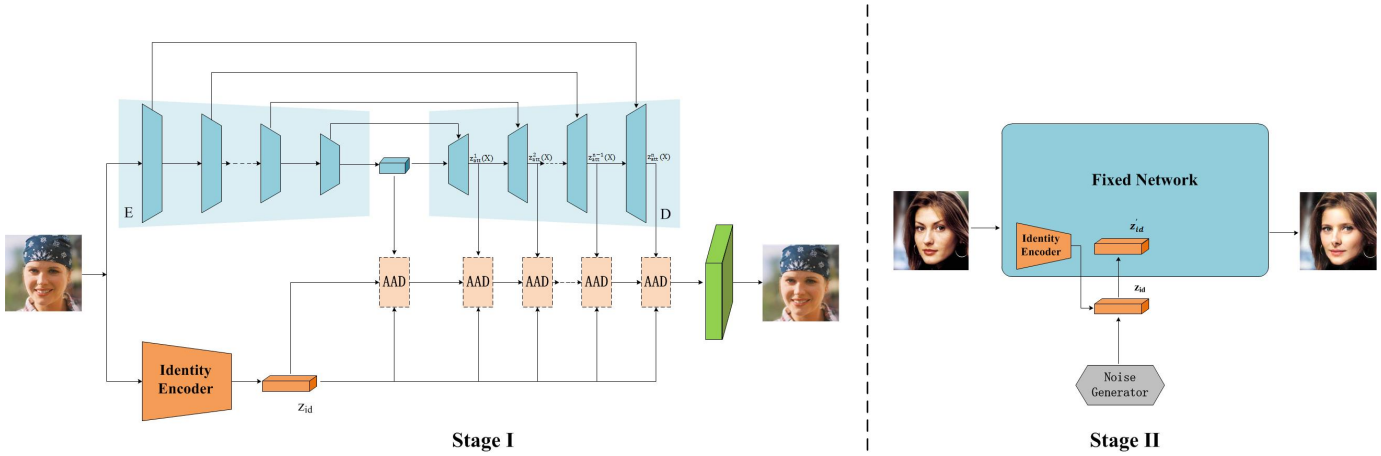

Fig. 1. Our approach involves two stages: (I) we train a GAN which is able to separate identity embedding as well as attribute embedding and then fuse them to obtain a face same as the original one. (II) we keep the trained neural network fixed, generate suitable noise, and add noise directly to the identity embedding. Here E means encoder and D means decoder.

that allow individuals to obtain visually similar face photos while hiding the identity information. In stage I, inspired by CNN's outstanding potential in latent space embedding disentanglement, we utilize a specially designed GAN without any other additional auxiliary structures, which can obtain the latent space embedding of a face photo individually for identity and attributes. Here the embedding of identity affects the verification tool to judge whether it is the same person, while the embedding of attributes guarantees the visual similarity. In stage II, because the GAN has been trained to learn how to obtain the disentangled identity embedding as well as attribute embedding, and to reconstruct a face from them, adjustable noise can be added directly to the latent identity embedding to break the above dilemma according to the user's demand.

The rest of this paper is organized as follows. Section II formulates the research problem and proposes our detailed two-stage approach. We evaluate our proposed method in Section III and draw some conclusions in Section IV.

II. Methods

A. Problem Formulation

Face de-identification problem can be formulated as a transformation function δ which maps a given face image X to a de-identified image $\hat{X}$, aiming to mislead the face verificators. The problem can be formulates as follows

$$\delta(X) = \hat{X} \tag{1}$$

$$s.t. : verificator\{X, \hat{X}\} < 0.5 \tag{2}$$

Mowever, $\hat{X}$ should look as similar as possible to the original face X considering image usability. Most current de-identified methods are designed for known-structure-verificators, here we perform a black box attack, and adopt the Google face verification API widely used in daily life as our verificator.

B. Two Stage Approach

We propose a novel face de-identified approach for identity obfuscation that combines a data-driven method with randomization. Our approach consists of two stages (see Fig. 1).

Stage I: In the first stage, we use a pretrained state-of-the-art face recognition model [13] as identity encoder. The identity embedding $z_{id}(X)$ is defined to be the last feature vector generated before the final FC layer. Face attributes, such as pose, expression, lighting and background, require more spatial information than identity. In order to preserve such details, we propose to represent the attribute embedding as multi-level feature maps. In specific, we feed the target image X into a U-Net-like structure, and then use the feature maps generated from the U-Net decoder as the attributes embedding. More formally, we denote

$$z_{att}(X) = \left\{z^1_{att}(X), z^2_{att}(X), \cdots, z^n_{att}(X)\right\} \tag{3}$$

where $z^k_{att}(X)$ represents the k-th level feature map from the U-Net decoder, n is the number of feature levels.

This attributes embedding network does not require any attribute annotations, it extracts the attributes using self-supervised training: we required that the generated de-identified face $\hat{X}$ and the original face X have the same attributes embedding. The loss function will be introduced in Equation 5.

In the end, we utilize a novel *Adaptive Attentional Denormalization* (AAD) layer designed by *Li et al.* [14] to integrate the effective regions of the identity embedding and the attributes embedding in an adaptive fashion.

Stage II: In the second stage, we keep the trained neural network fixed, generate suitable Gaussian noise according to the user's demand, and add the noise directly to the identity embedding.

Thanks to the great latent space embedding disentanglement ability of CNN, the generated face is visually similar to the original face, but can be misleading for face verification.

C. Training Process

Our framework is based on the AEI-Net model [14], however, the number of attribute embedding here is changed to $n = 6$ for training convenience. Consequently the corresponding network structure is adjusted.

We utilize adversarial training for this network. Let L_{adv} be the adversarial loss for making $\hat{X}$ realistic. It is implemented as a multi-scale discriminator [15] on the downsampled output images. An identity preservation loss is used to preserve the identity of the source. It is formulated as

$$L_{id} = 1 - \cos(z_{id}(\hat{X}), z_{id}(X)), \tag{4}$$

where $\cos(\cdot, \cdot)$ represents the cosine similarity of two vectors. We also use the attributes preservation loss, which is formulated as

$$L_{att} = \frac{1}{2} \sum_{k=1}^{n} \left\| z_{att}^{k}(\hat{X}) - z_{att}^{k}(X) \right\|_2^2, \tag{5}$$

The reconstruction loss as pixel level L-2 distances between the target image $\hat{X}$ and X

$$L_{rec} = \frac{1}{2} \left\| \hat{X} - X \right\|_2^2, \tag{6}$$

The full objective to train our network in the first stage is a weighted sum of above losses as

$$L_{total} = L_{adv} + \lambda_{att} L_{att} + \lambda_{id} L_{id} + \lambda_{rec} L_{rec}, \tag{7}$$

The above is a description of our idea, but in order to use visualization tools to judge the training effect and make appropriate adjustments in time, a little tricks are used here. In the first stage, we extract identity embedding and attribute embedding from two faces randomly sampled from the dataset and then fuse together, so that we can judge the training situation of each part of the network according to the real-time display. The reconstruction loss should be set to $L_{rec} = 0$ when the two faces are different.

III. Experiments

A. Implementation Detail

In stage I, to train our network, we use the CelebA-HQ dataset, which contains 30K high-resolution celebrity images. We randomly select 27K images for training and 3K for testing. All the images are resized to 256×256 in our experiments. We use the Adam optimizer with momentum parameters $\beta_1 = 0, \beta_2 = 0.999$. The learning rate is set to 0.0004. The parameters in Eq. (7) are fixed at $\lambda_{att} = \lambda_{rec} = 10, \lambda_{id} = 5$. In stage II, we first generate the required Gaussian noise of the same dimension as the identity embedding. Then we add the noise directly on the identity embedding while keeping the rest of the network unchanged to generate the de-identified face. All the evaluation results presented in this section are based on the test set to ensure fairness.

B. Comparison with Previous Methods

Now we compare the performance of our de-identified methods with both the traditional obfuscation methods and state-of-the-art methods: masking, mosaic, *Xue et al.* [6] and AnonymousNet[9]. The Google face verification API can return a confidence score between every generated face and its original face. The rule of judging whether the de-identified method successes refers to Equation 2.

Qualitative Results: To study the effect of noise on the generated image, we add different amounts of noise to the identity embedding in stage II. As shown in Fig. 2, when the added noise increases, the generated de-identified face looks less and less similar as the original face, so we can provide users with results according to their demands. It is worth noting that even when the noise is large, the output has little distortion and the result is still a face photo.

Fig. 2. The face images generated by adding different amounts of Laplace noise. For each row, the amount of noise added gradually increases from left to right; for each column, the amount of noise added is the same. For the parameters of noise, loc denotes the center of the Gaussian distribution, while scale denotes the width of the distribution. The score below each face image is the confidence score.

Then we compare the performance of ours with the methods mentioned before. As shown in Fig. 3, the Masking and Mosaic methods can hide the privacy information from human eyes, while human eyes can notice these changes easily. Moreover, these techniques are vulnerable for the verificator can still make judgement properly in most cases. *Xue et al.* obtained visually best result, however, due to it was designed exactly for Mask R-CNN, it can interfere well with Mask R-CNN's detection and judgment, but performs poor on unseen

verification tools. AnonymousNet can generate de-identified faces successfully in most cases, however, it suffers from distortion and artifacts in almost all samples. Our method not only ensures the removal of identity by deceiving the verificator, but also produces more visually pleasing results, which guarantees the data usability. We achieve higher fidelity by well preserving the face structure of the original image, while faithfully respecting its illumination and image resolution.

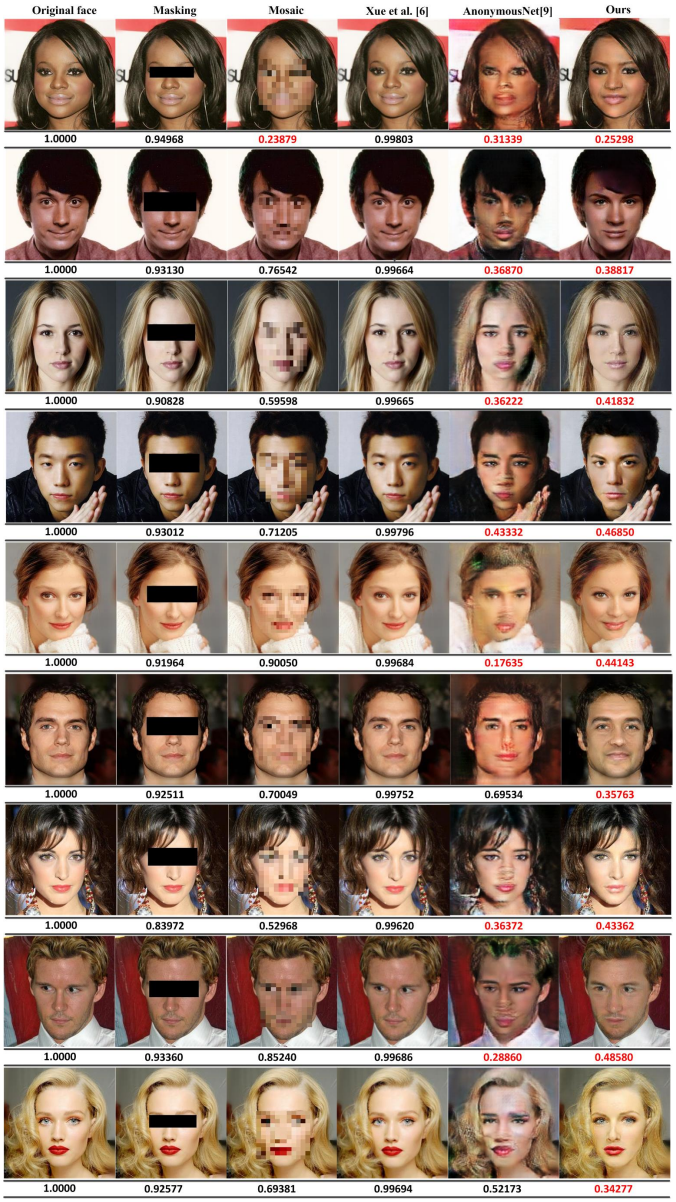

Fig. 3. Comparison with previous methods. Besides the advantages in de-identification, our results preserve the face structure of the original image, produce more visual-pleasing results and guarantee the image usability. The red score indicates successful de-identification while the black score indicates failure

Quantitative Comparison: We choose PSNR and SSIM as the metric for quantitative evaluation. Either a higher PSNR score or a higher SSIM score means a smaller distortion of the image. Table I lists the comparisons of our method with others. As we can see, among successful hiding human identity methods, our method performs better than AnonymousNet, and is comparable with traditional methods. Being aware that these model-based metrics fail to capture many nuances of human perception, we list the results here for reference purpose.

TABLE I
IMAGE QUALITY COMPARISON UNDER DIFFERENT METRICS

	Masking	Mosaic	*Xue et al.* [6]	AnonymousNet[9]	Ours
PSNR	16.0052	25.7771	25.1411	20.5089	24.6852
SSIM	0.9066	0.8868	0.7974	0.6564	0.8766

IV. Conclusions

In this paper, we are motivated by providing fine-grained control over identity information leakage in face images. Towards this goal, we present a novel two-stage method to generate visually pleasing face photos while hiding the identity information according to a single input. Meanwhile, we try to preserve visual similarity as much as possible to retain data usability by adding noise directly on the identity embedding. In the experiments, our results show that identity information in face images can be well protected while generating de-identified images of much higher visual realism. Therefore it can protect the image privacy under the premise of keeping the utility to the greatest extent. Moreover, our method adapts well to the verificator of unknown structure, which further improves the practical value in our real life.

References

[1] T. Orekondy, B. Schiele, and M. Fritz, "Towards a visual privacy advisor: Understanding and predicting privacy risks in images," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 3706–3715, IEEE Computer Society, 2017.

[2] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," in *European Conference on Computer Vision*, pp. 19–35, Springer, 2016.

[3] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *CoRR*, vol. abs/1609.00408, 2016.

[4] Y. Li, N. Vishwamitra, B. P. Knijnenburg, H. Hu, and K. Caine, "Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1343–1351, IEEE Computer Society, 2017.

[5] B. Liu, J. Xiong, Y. Wu, M. Ding, and C. M. Wu, "Protecting multimedia privacy from both humans and ai," in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, IEEE, 2019.

[6] H. Xue, B. Liu, M. Ding, L. Song, and T. Zhu, "Hiding private information in images from AI," in *2020 IEEE International Conference on Communications (ICC): Communication and Information Systems Security Symposium (IEEE ICC'20 - CISS Symposium)*, (Dublin, Ireland), June 2020.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, and D. Warde-Farley, "Generative adversarial nets in advances in neural information processing systems (nips)," 2014.

[8] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-gan for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, 2019.

[9] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[10] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2414–2423, IEEE Computer Society, 2016.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[12] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, pp. 1988–1996, 2014.
[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.
[14] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.
[15] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.