A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Towards Generating High Definition Face Images from Deep Templates

Xingbo Dong[1], Zhe Jin[2], Zhenhua Guo[3], Andrew Beng Jin Teoh[4]

Abstract: Face recognition based on deep convolutional neural networks (CNN) has manifested superior accuracy. Despite the high discriminability of deep features generated by CNN, the vulnerability of the deep feature is often overlooked and leads to the security and privacy concerns, particularly the risks of reconstructing face images from the deep templates. In this paper, we propose a method to generate high definition (HD) face images from deep features. To be specific, the deep features extracted from CNN are mapped to the input (latent vector) of the pre-trained StyleGAN2 using a regression model. Subsequently, HD face images can be generated based on the latent vector by the pre-trained StyleGAN2 model. To evaluate our method, we derived the face features from the generated HD face images and compared them against the bona fide face features. In the sense of face image reconstruction, our method is simple, yet the experimental results suggest the effectiveness, which achieves an attack performance as high as SAR=46.08% (18.30%) @ FAR=0.1 threshold under type-I (type-II) attack settings. Besides, experiment results also indicate that 50.7% of the generated HD face images can pass one commercial off-the-shelf (COTS) liveness detection.

Keywords: Face template security, face image reconstruction, deep templates.

1 Introduction

The recent thriving of deep learning technology has succeeded in numerous computer vision applications such as face recognition (FR). In fact, the deep learning enabled approach has become a de-facto standard for face recognition due to the superior recognition performance. In general, a deep learning-based FR system is composed of three main components, i.e., pre-processing, convolution neural network based feature extractor, and a matcher. Despite enjoying decent performance and convenience, security and privacy concerns on FR systems rise among the public because of the inherent linkage between the face data and the owner identity. For example, the disclosure of face data may expose the private and sensitive information of the user (e.g., race, age, gender). Moreover, face templates could be inverted, hence face images can be generated to gain illegal access to the system.

A number of work have been done on the restoration of the face template to the face image [MJ13, Ma18]. One of the latest works, namely a neighborly de-convolutional neural

[1] School of Information Technology, Monash University, Malaysia Campus, Malaysia, xingbo.dong@monash.edu
[2] School of Information Technology, Monash University, Malaysia Campus, Malaysia, jin.zhe@monash.edu
[3] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, zhenhua.guo@sz.tsinghua.edu.cn
[4] School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, bjteoh@yonsei.ac.kr

network (NbNet) is proposed to reconstruct the face image from its deep feature counterpart [Ma18]. The generated face images can achieve an attack success rate of 95.20% (58.05%) on LFW dataset [Hu08] under type-I (type-II) attacks with a false accept rate of 0.1%. The type-I attack here refers to comparing the generated fake image against the face image from which the template was extracted. In contrast, type-II refers to comparing the generated fake image against the bona fide[3] face images of the same subject that were not used for template creation.

However, the generated face images in [Ma18] are in low-resolution and leaves over the clue of synthetical artifacts, which is easy to be detected by the image based face liveness detection. As opposed to the low-resolution face images, high definition (HD) face images aid adversaries to perform the attack easily and efficiently, e.g., construct a 3D face mask from an HD face image. In this paper, we focus on generating HD face images based on the deep features[4] and attempt to utilize the generated HD face images to access a targeted FR system by launching type-I and type-II attacks. The proposed method to generate an HD face image based on the deep features can be regarded as a partial invertible scheme without precisely recovering the original face data, which may lead to privacy leakage and security compromises. Our work may shed light on the biometric deployment to meet the privacy-preserving, and security policies, such as the EU General Data Protection Regulation (GDPR) [Co18].

In summary, this paper makes the following contributions:

1. By learning a mapping network between the latent vector space of a pre-trained StyleGAN2 model and the feature vector space of a pre-trained CNN extractor, a simple but effective method to generate HD face images from deep features is established.
2. Attacks to compromise the security of the face recognition systems are simulated. Specifically, the CNN-based face feature extractor is regarded as a black-box in the attacks. The generated HD face images are used to compare with the original face features to simulate the compromising of security of the face recognition systems.
3. The proposed method is also evaluated with a Commercial-Off-The-Shelf (COTS) face liveness detector. It shows that the generated HD face images manage to fool the COTS detector.

In the remainder of this paper, some existing related works will be reviewed in Section 2, and the detailed methodology will be discussed in Section 3. The experiments and results are shown in Section 4. Finally, the conclusion is drawn in Section 5.

[3] We refer to Bona fide as genuine face images.
[4] Deep features and deep templates are used interchangeably.

2 Related works

Face image generation can be traced back to face image reconstruction [RdJG98, Tr99, Tr06]. The application of such face image reconstruction can facilitate the witnesses in a crime scene. For example, as presented in [Tr06], generated face images are shown to the witness, and the candidate face images are selected by the witness. The selected candidate face images are further evaluated based on an optimization algorithm by narrowing the eigenface coefficient space iteratively and then generated face images are computed and shown to the user again. To achieve this task, the user's interactive input is always required in such a system.

In [Ad03, FLY14], hill-climbing is utilized to generate a synthetic face image from a corresponding real-valued template. Specifically, a random face image is initiated firstly; then, the face image is perturbed iteratively based on the matching score between the current iteration and the previous iteration-based synthetic face image' features. The iteration is ended when the matching score decreases to a decision threshold. The corresponding synthetic face image is used as the final output. In [MJ13], radial basis function (RBF) regression in the face eigenspace is adopted to reconstruct visually realistic face images from the local pattern features. In [MSK07], a scheme to reconstruct face images from match scores is developed. An affine transformation is utilized to approximate the behavior of the face recognition system. A similarity score matrix is generated firstly based on the target face recognition system, and an affine space is subsequently learned based on the similarity matrix. Given the distances of the targeted subject's template, the template is embedded in the affine space; an affine transformation is applied to retrieve the original template.

In the era of deep learning, FR systems based on deep learning models have been widely deployed. Reconstructing face images from the deep features draws the attention of the public due to the privacy and security concerns. To generate face images from deep templates, two main branches have been proposed in the literature, i.e., white-box based (feature extractor model is known) and black-box based (feature extractor model is unknown) approaches. [ZS16, Co17] are typical white-box based approaches while [Ma18] is a black-box approach.

[ZS16] proposed a method to invert FaceNet face embedding [SKP15] to realistic-looking face images based on convolutional neural networks. Specifically, face image reconstruction is formulated as a minimization problem that attempts to minimize the template difference between original and reconstructed images. However, a regularization function constructed using the intermediate nodes of the target extractor model network is required in the proposed scheme. Hence the detailed parameter of the target template extractor should be known. In reality, however, the extractor model may not usually be available. In [Co17], a method to synthesize a frontal, neutral expression face image from the FaceNet feature [SKP15] is proposed. Firstly, the landmarks and textures of face images are estimated by off-the-shelf landmark detection tools and a warping technique. Face images are then generated based on differentiable image warping by combining landmarks and textures information. In the implementation, however, the last convolution layer instead of the

final output of a pre-trained FaceNet model is used; hence the parameters of the extractor model, i.e., FaceNet, should be known.

In [Ma18], a neighborly de-convolutional neural network (NbNet) is designed to reconstruct face images from their deep templates. Unlike the aforementioned models, the knowledge of the target subject and the deep network are not required. Specifically, the NbNet is a cascade of multiple stacked de-convolution blocks and a convolution block. Unlike the conventional convolution operation, de-convolution operations can up-sample the input data to produce a larger output feature map. Subsequently, a convolution operation is applied on the output of the de-convolution output to generate the output face images. In [Ma18], GAN synthesized face images, and two augmented benchmark face datasets are used to train the model. The system is evaluated with type-I and type-II attack settings.

Although, a variety of approaches to reconstruct face images from deep features are reported. The face images generated from the aforementioned approaches are not in HD resolution. For example, the output image size in [Ma18] is 160×160, and the size in [Co17, ZS16] is 224×224. Such low resolution may not meet the requirement of some applications, for example, face liveness detection.

On the other hand, a number of techniques that generate face images based on adversarial models had been proposed. Among various techniques, StyleGAN2 [KLA19, Ka20] is one of the most popular methods to generate high-resolution and realistic face images. In StyleGAN2, a mapping network is used to map points in latent space to an intermediate latent space, then the intermediate latent space is utilized to control the style in the generator model.

In this paper, we propose an alternative way to achieve template inverting tasks by incorporating the StyleGAN2's capabilities to generate face images from the deep templates, and show that generating HD face images may threaten FR systems, especially liveness detection.

3 Generating HD Face Images from deep features

A method to generate HD face images based on deep features is presented in this section. We firstly assume that the stored template $v = f(x) \in \mathscr{V}$ is known to the adversary, $\mathscr{V}$ denotes the feature space. We also assume that the adversary can generate unlimited input-output data pairs of the deep feature extractor. However, the deep feature extractor is regarded as a black-box that is not necessarily known to the adversary. By learning a mapping between latent vector space of the StyleGAN2 and the feature vector space of the face feature extraction model, latent code of the corresponding compromised template can be predicted. Next, a pre-trained StyleGAN2 model [Ka20] is utilized to generate HD face images by exploiting the information originating from the deep features. Next, those generated fake images are used to access (attack) a target system illegally. An overview of the method is shown in Fig. 1.

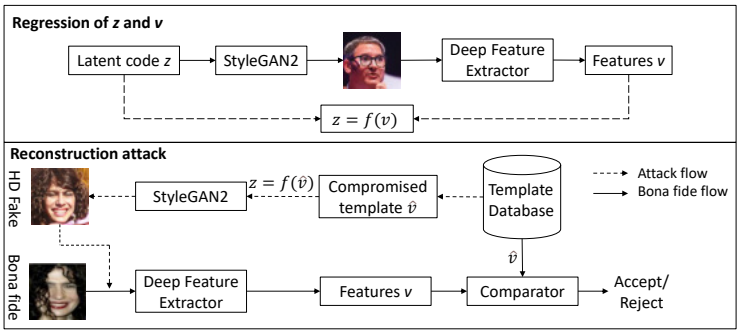


Fig. 1: An overview of the proposed method.

We expect that the fake face images generated based on deep features should preserve some characteristics from the original face image, e.g., gender and age. Besides, we also expect that the generated fake face images can also be utilized to gain illegal access to the system by launching type-I and type-II attacks. To achieve this task, the input of the StyleGAN2, i.e., the latent code, should be determined first. The problem now can be defined as given a deep features v, how to determine the corresponding latent code z.

To solve above problem, a regression model based on neural networks is established to learn the mapping between latent space and feature space and denoted as the mapping network. The mapping network has two hidden layers with 1024 nodes having a relu activation function. The input and output size of the network are both 512, while the linear activation function is used for the output layer as this is a regression task. Finally, the model is trained based on the MSE loss by Adam optimizer. Given an input face features v, the model is expected to output the corresponding latent code z.

To obtain the training data to train the regression model, 3.2 million $< v_i, z_i >$ pairs are generated based on the adopted pre-built models. Specifically, the random latent code vector $z_i \in \mathbb{R}^{512}$ is generated pirorly, then the latent code z_i is supplied as an input to the pre-trained StyleGAN2 model, and a fake face image can be generated. Subsequently, face feature v_i is extracted from the generated image by the InsightFace model. It is worth highlighting that no dataset is needed to train the regression model, as we directly utilize the pre-trained model to generate the training samples.

It is worth highlighting that the proposed method to generate face images from features enjoys advantages compared with existing methods:

1. HD face images are generated in our scheme. Compared with the existing works such as 160×160 in [Ma18] and 224×224 in [Co17, ZS16], the generated face images in this paper is in 1024×1024 resolution. HD face images provide extra advantages to perform the attack, such as liveness detection.
2. Our pipeline is concisely simple and efficient, which only needs to train a regression model. Besides, no extra training dataset is needed. Simultaneously, the attack can still be feasible.

Tab. 1: SARs of type-I and type-II attacks on LFW.

FAR	Normal TAR	Threshold	Type I	Type II
0.0%	91.77%	0.4183	0.49%	0.10%
0.1%	93.80%	0.4008	1.42%	0.46%
1.0%	97.40%	0.3669	10.11%	3.13%
10.0%	98.93%	0.3337	46.08%	18.30%

4 Experiments and Results

In our experiment, a pre-trained StyleGAN2 model[5] is adopted. The pre-trained model is trained on FFHQ dataset [KLA19] at 1024×1024. The 512-D face features were extracted by the InsightFace (ArcFace) with ResNet-100 backbone [De18] [6].

To evaluate the performance of the proposed method, the Labeled Faces in the Wild (LFW) [Hu08] face dataset is adopted in this experiment. The face features of LFW bona fide face images are extracted. Next, the face features are fed into the regression model to compute the corresponding latent code vectors. Then the fake HD face images are generated by the pre-trained StyleGAN2.

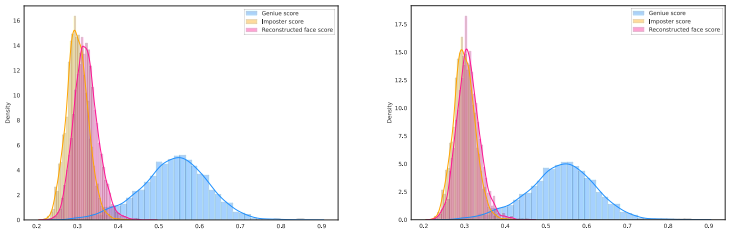
To simulate the attack, the generated fake face images are then directly used as the input of the feature extractor to extract the features. Finally, features are compared with the stored template to compute the similarity score.

We quantitatively evaluated the security of the deep features under type-I and type-II attacks. Specifically, the official LFW verification protocol[7] is adopted in this paper to compute true accept rate (TAR) at different false accept rate (FAR) on the deep features. To distinguish with the TAR in a normal situation, the TAR corresponding to specific FAR under attack situations is denoted as Success Attack Rate (SARs), and higher SAR means a high risk of compromising. The results are shown in Table 1.

From the table 1, we observe that the generated face images can achieve relatively high SARs under type-I attack settings. The SAR can reach 46% under the threshold at FAR=10%, and the attack SAR is nearly 5 times higher than the false accept attack rate (FAR=10%) by attempting the access with a random imposter sample, which implies high risks of adversary attacks. Under FAR=1%, the attack SAR under type-I can still achieve 10.11%. On the other hand, the type II attack shows weaker performance than type-I, as the attack SAR can only reach 18.30% under the FAR=10% threshold. However, the risks still persist under this setting.

To show the difference between the proposed attack and the false accept attack, the comparing scores (similarity) distribution between genuine pairs, imposter pairs, and attack pairs are shown in Fig. 2. It is seen that the generated HD face images can generate higher similarity scores than a random imposter sample. This suggests that the proposed method

[5] https://nvlabs-fi-cdn.nvidia.com/StyleGAN2/networks/StyleGAN2-ffhq-config-f.pkl
[6] https://www.dropbox.com/s/tj96fsm6t6rq8ye/model-r100-arcface-ms1m-refine-v2.zip
[7] http://vis-www.cs.umass.edu/lfw/#views



(a) Type-I (b) Type-II

Fig. 2: Scores (similarity) distribution between genuine pairs, imposter pairs and attack pairs.

Tab. 2: Comparison with the state-of-the-art methods

	Deep feature extractor	Resolution	Liveness pass rate
Mai et al. [Ma18] (VGG-NbA-P)	Facenet	160×160	1.77%
Ours	InsightFace	1024×1024	50.70%

is not a false acceptance attack and more vulnerable in the practical biometric systems deployment.

Examples of the generated HD face images based on randomly selected subjects in LFW are shown in Fig. 3. We can find that the succeeded fake face instances show high similarity visually compared with the corresponding bona fide images. Failure cases are also shown in Fig. 3 (c). The results suggest that the proposed method could be a real threat to practical FR systems.

To further validate the advantage of the HD face images, a COTS liveness detection cloud computing API is used to evaluate the generated face images. The COTS returns three suggestions for each image, i.e., block, review, and pass. If the image is detected as block class, then this face image will be regarded as non-live. The detailed results are shown in Fig. 4.

As shown in Fig. 4, 50.7% of our generated HD face images manage to fool the face liveness detector, while 60.4% images from [Ma18] under VGG-NbA-P setting are blocked. This is because the generated face images from [Ma18] are in low-resolution, and also contain artifacts. Examples of blocked and passed face images generated by our model can be found in Fig. 5.

Table 2 shows a comparison with one of the current state-of-the-art schemes. [Ma18] shows superior performance due to the specific designed neighborly de-convolutional neural network (NbNet). But surprisingly, the combination of StyleGAN2 and a simple regression can still achieve 10.11% SAR at the FAR=1%.

(a) Success attack examples at FAR=0.1% threshold.

(b) Success attack examples at FAR=1% threshold.

(c) Failed attack examples at FAR=1% threshold.

Fig. 3: Generated HD face images from the randomly selected subjects in LFW. (a) and (b) shows the success attack examples at the FAR=0.1% and FAR=1% threshold, respectively. (c) shows the failure attack examples at the FAR=0.1% threshold (Best view in color and zoom in).

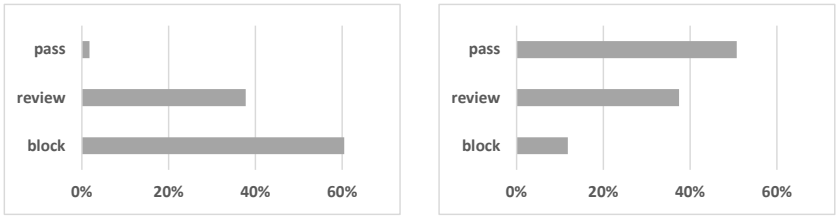
5 Conclusion

In this paper, a method to generate HD face images from the deep features has been presented. By finding a mapping between the latent space and the feature space, the method allows HD face images generation from deep face features with StyleGAN2. The forged HD face images do not need to resemble exactly the bona fide face images. However, it could be exploited by the adversary to gain illegal access to the face recognition systems.

We also simulate type-I and type-II attacks on the LFW dataset, and the quantitative results show that the proposed method can achieve comparable performance. Compared with state-of-the-artwork in [Ma18], our method is simple, and higher resolution images can be attained.

The generated HD face images are also evaluated by a COTS liveness detection API, and it shows that 50% of samples can pass the liveness detection system. This indicates that the current usage of COTS photo-based liveness detection API is at risk and still needs to be improved.

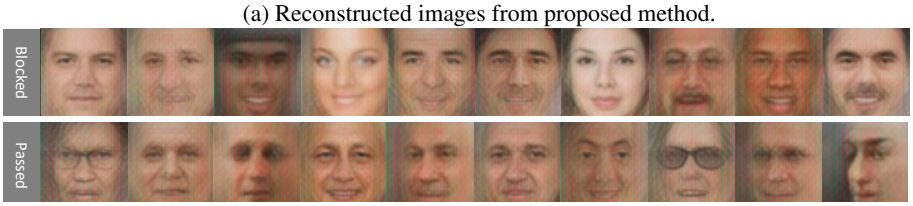
It is interesting to extend the proposed approach into the investigation of existing biometric template protection algorithms. For example, by finding a mapping between the protected (or transformed) template space and feature space, is it possible to generate HD face im-



(a) Mai et al. [Ma18] (b) Ours

Fig. 4: Liveness detection results based on COTS .

(a) Reconstructed images from proposed method.



(b) Reconstructed images from [Ma18]. (Source images provided by Mai.)

Fig. 5: Blocked and passed examples.

ages to compromise the security? On the other hand, StyleGAN2 model shows impressive performance in generating HD face images, but the modeling capability of StyleGAN2 model is limited in our work as it is pre-trained. The attack SAR could be improved by fine-tuning the model with specific training data. In addition, utilizing the reconstructed images to attack different face recognition systems is also an interesting future investigation.

Acknowledgement

This work was supported by grants from Ministry of Higher Education (MOHE) Malaysia through Fundamental Research Grant Scheme (FRGS/1/2018/ICT02/ MUSM/03/3). The authors would like to thank Dr. Mai Guangcan for his gratitude in offering the reconstructed face images.

References

[Ad03] Adler, Andy: Sample images can be independently restored from face recognition templates. In: CCECE 2003-Canadian Conference on Electrical and Computer Engineering.

Toward a Caring and Humane Technology (Cat. No. 03CH37436). volume 2. IEEE, pp. 1163–1166, 2003.

[Co17] Cole, Forrester; Belanger, David; Krishnan, Dilip; Sarna, Aaron; Mosseri, Inbar; Freeman, William T: Synthesizing normalized faces from facial identity features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3703–3712, 2017.

[Co18] Commission, European: , 2018 reform of EU data protection rules, 2018.

[De18] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: Arcface: Additive angular margin loss for deep face recognition. arXiv preprint arXiv:1801.07698, 2018.

[FLY14] Feng, Yi C; Lim, Meng-Hui; Yuen, Pong C: Masquerade attack on transform-based binary-template protection based on perceptron learning. Pattern Recognition, 47(9):3019–3033, 2014.

[Hu08] Huang, Gary B; Mattar, Marwan; Berg, Tamara; Learned-Miller, Eric: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, alignment, and recognition. 2008.

[Ka20] Karras, Tero; Laine, Samuli; Aittala, Miika; Hellsten, Janne; Lehtinen, Jaakko; Aila, Timo: Analyzing and Improving the Image Quality of StyleGAN. In: Proc. CVPR. 2020.

[KLA19] Karras, Tero; Laine, Samuli; Aila, Timo: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410, 2019.

[Ma18] Mai, Guangcan; Cao, Kai; Yuen, Pong C; Jain, Anil K: On the reconstruction of face images from deep face templates. IEEE transactions on pattern analysis and machine intelligence, 41(5):1188–1202, 2018.

[MJ13] Mignon, Alexis; Jurie, Frédéric: Reconstructing faces from their signatures using RBF regression. In: British Machine Vision Conference 2013. pp. 103–1, 2013.

[MSK07] Mohanty, Pranab; Sarkar, Sudeep; Kasturi, Rangachar: From scores to face templates: a model-based approach. IEEE transactions on pattern analysis and machine intelligence, 29(12):2065–2078, 2007.

[RdJG98] Rosenthal, Y; de Jager, G; Greene, J: A computerised face recall system using eigenfaces. In: Proceedings of the Eighth Annual South African Workshop on Pattern Recognition. pp. 53–57, 1998.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823, 2015.

[Tr99] Tredoux, Colin; Rosenthal, Yon; da Costa, Lisa; Nuenz, D: Face reconstruction using a configural, eigenface-based composite system. SARMAC III, 1999.

[Tr06] Tredoux, Colin; Nunez, David; Oxtoby, Oliver; Prag, Bhavesh: An evaluation of ID: an eigenface based construction system: reviewed article. South African Computer Journal, 2006(37):90–97, 2006.

[ZS16] Zhmoginov, Andrey; Sandler, Mark: Inverting face embeddings with convolutional neural networks. arXiv preprint arXiv:1606.04189, 2016.