Model-Free Template Reconstruction Attack with Feature Converter

A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

Model-Free Template Reconstruction Attack with Feature Converter

Muku Akasaka [1] Yuya Sato [1] Soshi Maeda [1] Masakatsu Nishigaki [1] Tetsushi Ohki [1]

Abstract: State-of-the-art template reconstruction attacks assume that an adversary has access to a part or whole of the functionality of a target model. However, in a practical scenario, rigid protection of the target system prevents them from gaining knowledge of the target model. In this paper, we propose a novel template reconstruction attack method utilizing a feature converter. The feature converter enables an adversary to reconstruct an image from a corresponding compromised template without knowledge about the target model. The proposed method was evaluated with qualitative and quantitative measures. We achieved the Successful Attack Rate(SAR) of 0.90 on Labeled Faces in the Wild Dataset(LFW) with compromised templates of only 1280 identities.

Keywords: Template reconstruction attack, face recognition, template security, model inversion.

1 Introduction

Biometric recognition systems have the advantage of recognition capability without the need to remember additional information. However, most biometric data are almost immutable for a person's lifetime, which can cause crucial privacy problems if such data are compromised. Therefore, the privacy of biometric data should be considered while maintaining its utility.

A template reconstruction attack is one of the typical privacy threats to a biometric system. A biometric template is a set of stored biometric features directly comparable to probe biometric features. The attack aims to reconstruct the biometric sample used for generating the biometric templates by exploiting the comparison results of the system, biometric features or biometric templates. In template reconstruction attacks, the adversary's objective is not only to spoof the system but also to know the appearance of the original image used to calculate the biometric template. The trial of image reconstruction from biometric templates starts with the basic concept of score-based hill-climbing[UJ04, Ad04]. Recent template reconstruction attack approaches, such as those of [Ma19, Du20], successfully reconstruct face images from biometric templates by training an inversion model with face images and face feature sets. However, in previous studies, it was assumed that an adversary has unlimited access to the feature extractor of the target model in both *white-box* and *black-box* settings. The feature extractor is a core function of biometric systems; therefore, they should be strictly protected by the administrator. In previous studies [Ma19], it was also assumed that feature extractors are publicly available via an SDK. However, these

[1] Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, JAPAN, {akasaka, sato, maeda}@sec.inf.shizuoka.ac.jp, {ohki, nisigaki}@inf.shizuoka.ac.jp

assumptions are not realistic in practical scenarios when template reconstruction attacks target commercially developed systems.

In contrast, we propose a method that enables an adversary to reconstruct images from compromised templates even when they do not have complete information about the feature extractor. In the method, the adversary can exploit limited instructive knowledge obtained in a practical scenario by enrolling a set of face images even in a strict *black-box* setting. We call such a *black-box* attack a *model-free* attack where an adversary cannot procure any target models for unlimited access. Specifically, we prepare a feature converter and a generic feature extractor different from the feature extractor in the system. The feature converter translates a biometric feature for the target feature extractor to that for the generic feature extractor. In most cases, the adversary can enroll biometric data in the target system even when access to the feature extractor is restricted. This helps the adversary train the feature converter using only a few pairs of pre-enrolled biometric samples and biometric features. The contributions of this work are summarized below.

- We propose a novel *black-box* attack called a **model-free** template reconstruction attack that utilizes a feature converter from an unknown private space to a known public space. This method can reconstruct highly accurate images under the realistic assumption that the adversary can pre-enroll a small number of biometric samples.
- We evaluate the proposed method from three perspectives, that is, the amount of information an adversary can have, the performance of the impersonation attack, and the quality of the reconstructed data, and show its effectiveness against existing methods.

2 Related work

The study of face template reconstruction attacks starts with the hill-climbing techniques proposed by Soutar et al. [So02][6]. They used the cosine similarity between biometric templates and biometric features extracted by synthesized images to create a reconstructed image that maximizes a comparison score. The methods can be easily defended by quantizing comparison scores. To deal with the countermeasure, Adler et al. [Ad03, Ad04] proposed to process the image while adding arbitrary eigenfaces to the image. Considering that the above-mentioned methods target models that produce shallow features, in recent studies, the attack to deep templates was attempted by exploiting the feature extractor in the system. One problem in the image reconstruction task for a deep template is that hill-climbing can be easily stacked in a minimal local value caused by the complexity of biometric templates: this may end in a random noise image. Zhmoginov et al. [ZS16], an early study of deep template reconstruction attack, assumes that the adversary can access the feature extractor in a white-box scenario, which is an essential premise for the gradient descent approach. Mai et al.[Ma19] grappled with the more challenging black-box scenario by creating enormous pairs of public face images and corresponding templates

[6] We cannot access this article today. An overview of this paper was shown in Adler et al. [Ad03].

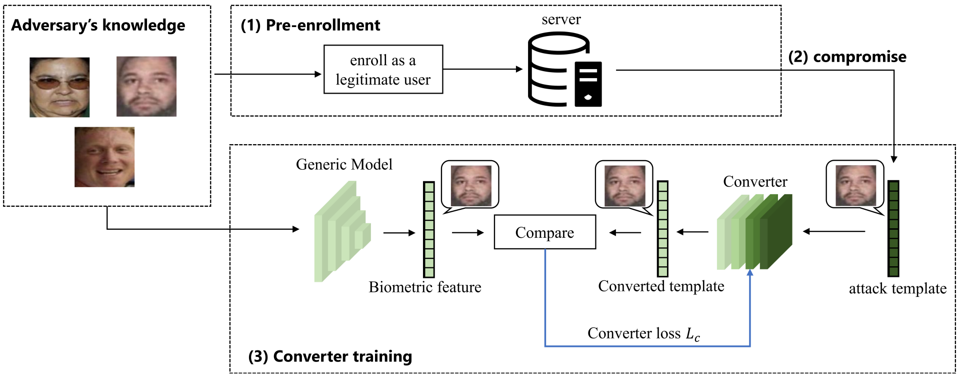


Fig. 1: Overview of converter training. Black lines indicate the data flow, and blue lines show the loss of variables. An adversary first enrolls a set of images as a legitimate user. The converter is trained to reduce the converter loss L_c, the difference between a compromised template and a biometric feature of the same identity calculated by a generic model.

with a target feature extractor to train the reconstruction template decoder, called NbNet. Xingbo et al. [Do21] utilized a GAN to improve the image quality of NbNet. Prior deep template reconstruction attacks have been based on the impractical assumption that the adversary has unlimited access to the feature extractor. Instead, we aim to satisfy the premise that the feature extractor cannot be compromised in a commercial service.

Model inversion attacks are similar to the template reconstruction attacks against machine learning models initially proposed by Fredrikson et al. [FJ15]. The significant difference between the two is that the model inversion attack aims to reconstruct the images included in the training dataset from the output of the multi-class classification result, while the template reconstruction attack reconstructs the image from the biometric template stored in the target system. Generative model inversion attacks, proposed by Zhang et al. [Zh20], exploited WGAN [ACB17] in the gradient descent method. We also leveraged WGAN in our proposed method, inspired by previous studies.

3 Proposed method

An overview of the proposed method is shown in Fig. 1 and Fig. 2. In this section, we discuss the threat model and the details of our proposed method.

3.1 Threat model

Traditional MI attacks assume multi-class classification, and the adversary capitalizes on the *white-box* knowledge of the target model to invert the training image. However, here, we expressly assume a feature-based face recognition system as the target system. A feature-based face recognition system takes a face image x as an input to the model and stores the biometric feature f extracted from the model as the biometric template t

associated with its corresponding identifier y. This study focuses on a *black-box* setting in which the adversary reconstructs the original face image x from compromised biometric template t without any knowledge of the target model T or the target feature extractor T_f.

Possible knowledge of the adversary. Let n be the number of enrolled users. Let $t_i \in t$ be the i-th biometric template made from biometric image x_i, i.e. $t = \{t_1 \cdots t_n\}$. t_i consists of an identifier y_i and biometric feature f_i, i.e. $t_i = \{y_i, f_i\}$. Even if they cannot access the target feature extractor T_f, an adversary can enroll m biometric templates $\hat{t} = \{\hat{t_1} \cdots \hat{t_m} \mid m \ll n\}$ associated with known biometric samples $\hat{x} = \{\hat{x_1} \cdots \hat{x_m} \mid m \ll n\}$ to the target system. When enrolled templates $\hat{t}$ are compromised from a target system, adversary can know the correspondence between the enrolled templates $\hat{t}$ and known biometric samples $\hat{x}$ on the target model without accessing the target feature extractor. Throughout this paper, we refer to such data intentionally enrolled by an adversary as attack samples $\hat{x}$, attack template $\hat{t}$, and attack feature $\hat{f}$. Note that, compared with the total number of enrolled biometric templates n, the number of attack samples m is infinitely small. The proposed method uses a small number of attack samples $\hat{x}$ and attack templates $\hat{t}$ to efficiently train feature converter C between different networks and reconstruct high-quality images.

3.2 Template reconstruction attack with feature converter

As shown in Fig. 2, our method uses four networks to reconstruct high-quality images from a limited number of biometric features, that is, a generator G, a discriminator D, a generic model A, and a converter C. More precisely, our method consists of three steps: (1) *Public model construction*, in which we train the generator, discriminator, and generic model from public datasets, (2) *Converter training*, in which we train the feature converter to convert the feature space from a target model to an attack model, and (3) *Template reconstruction*, in which we make use of the generator obtained from the first step and the converter from the second step and solve an optimization problem to reconstruct the original biometric sample from a compromised template.

For the (1) *Public model construction* step, the generator G and the discriminator D are trained on an available public dataset using a Wasserstein GAN [ACB17]. That is, we minimize the following training loss.

$$\min_G \max_D \mathcal{L}_{wgan}(G,D) = E_x[D(x)] - E_z[D(G(x))] \tag{1}$$

In addition, the adversary trains a generic model A on a publicly available dataset. The generic model A can prepare a face recognition model independent of the target model T.

For the (2) *Converter training* step, once the adversary obtains attack templates, the attack proceeds to train phase of a converter C. Fig. 1 shows the data flow of this step. The converter $C : \mathcal{F}_T \mapsto \mathcal{F}_A$ converts template in the target model's feature space $\mathcal{F}_T$ to a generic model's feature space $\mathcal{F}_A$ by reducing their cosine similarity. Let A_f denote the feature extractor of a generic model A, and $\mathrm{cossim}(\cdot,\cdot)$ denote the cosine similarity between two

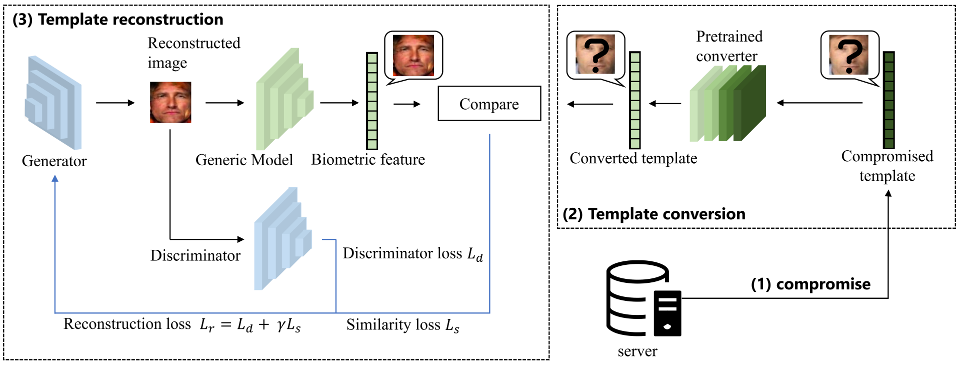


Fig. 2: Overview of template reconstruction attack phase. Black lines indicate the data flow, and blue lines show the loss of variables. An adversary aims to reconstruct an image from a compromised template of unknown identity. They search for an ideal latent vector of the generator navigated by the reconstruction loss.

biometric features. The converter loss $\mathcal{L}_C$ can thus be expressed as

$$\mathcal{L}_C = \sum_{i}^{m} \left[1 - \mathrm{cossim}(C(\hat{f}_i), A_f(\hat{x}_i)\right]. \tag{2}$$

In the (3) *Template reconstruction* step, we introduce a template reconstruction attack by a gradient descent algorithm. Fig. 2 shows the data flow of this step. An image can be reconstructed by searching the latent vector z to maximize the cosine similarity with a particular feature f. The reconstructed image can be explored by minimizing the loss function below, searching for the optimal z.

$$z = \underset{z}{\mathrm{argmin}} \left[-D(G(z)) - \gamma \left(\mathrm{cossim}(T_f(G(z)), f) \right) \right] \tag{3}$$

where γ is a weight to control the balance between the reality and resemblance of the reconstructed face image. However, in the *black-box* setting, the target model T in Equation (3) is inaccessible to the adversary. $C(f)$ in Equation (3) can replace f by template conversion with the pretrained feature converter C. Therefore, we transform Equation (3) using the pretrained converter as follows:

$$z = \underset{z}{\mathrm{argmin}} \left[-D(G(z)) - \gamma \left(\mathrm{cossim}(A_f(G(z)), C(f)) \right) \right]. \tag{4}$$

where the reconstruction loss $\mathcal{L}_r = -D(G(z)) - \gamma(cossim(A_f(G(z)), C(f)))$ consists of the discriminator loss $\mathcal{L}_d = -D(G(z))$ to encourage the reconstructed image to be a realistic image and the similarity loss $L_c = cossim(A_f(G(z)), C(f))$ to help it resemble the original image. Equation (4) enables an adversary to reconstruct the original face image without accessing the target model because both $A_f(G(z))$ and $C(f)$ belong to feature space $\mathcal{F}_A$. In our proposed method, the feature converter C consists of four fully convolutional layers of 512 neurons, alternating with three ReLU activation layers.

4 Experiment

We quantitively evaluated our method from the aspect of the reconstruction performance and the quality of the reconstructed images. We assumed a *model-free* setting where the adversary's knowledge is restricted to the pre-enrollment of certain amount of images.

4.1 Experimental setup

Datasets. We evaluate our method using four datasets: (1) CelebFaces Attributes Dataset [Li15] (`CelebA`), (2) CASIA-WebFace Dataset [Yi14] (`CASIA-WebFace`), (3) Labeled Faces in the Wild Dataset [Hu12] (`LFW`), and (4) IARPA Janus Benchmark-C Dataset [Ma18] (`IJB-C`). `IJB-C` was cropped by an official crop script, and images smaller than 160 × 160 pixels were screened out. All the other face images were aligned using MTCNN [Zh16] to crop them to 160 × 160 pixels.

Models. For a fair experiment, we leveraged two models of different architectures with different training datasets as the private target model T and the public generic model A. We used open-source models pretrained with VGGFace2 [PVZ15][7] and MS1Mv2 [De19][8] for T and A, respectively. In addition, we prepared public Generator G and public Discriminator D trained with all images of `CelebA`.

Protocol. For the *converter training*, we randomly sampled the exact number of attack samples $\hat{x}$ from each m user of the `CASIA-WebFace` dataset. Then, the corresponding attack features $\hat{f}$ were extracted from attack samples and used to train converter C. We used the Adam [KB15] optimizer with a batch size of 64 and a learning rate of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The training stopped when the update of loss stopped to avoid overfitting.

For the *template reconstruction* step, we sampled 300 samples from `CASIA-WebFace`, `LFW`, and `IJB-C` as evaluation templates $t = (y, f)$. Then we conduct the *template reconstruction* step with the converted feature $C(f)$ in the inference phase. The inference started with 256 independent 100-dimensional z split into two batches, and the image outputting the feature with the highest cosine similarity to the converted feature was chosen as the reconstructed image after 100 epochs. Then, we extracted feature from the reconstructed image using the target model and calculated matching score with the evaluation template. The search for the optimal image with the GAN was performed by Adam with a batch size of 128, a learning rate of 0.035, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We experimentally found $\lambda = 100$ to be the optimal weight for successful reconstruction.

Metrics. We employed the Inception Score [Sa16] and Successful Attack Rate(SAR) as a measure of image quality and a criterion of the impersonation performance, respectively. The Inception Score assesses image quality in the context of how well a multiclass classification model can identify the images. The SAR can be calculated by the False Accept

[7] facenet-pytorch: `https://github.com/timesler/facenet-pytorch`

[8] magface-pytorch: `https://github.com/IrvingMeng/MagFace`

Tab. 1: SAR of each model. Attack template for training the converter are extracted by CASIA-WebFace, and evaluation templates for the calculation of SAR are extracted by CASIA-WebFace, LFW, and IJB-C

Algorithm	Attack template size	SAR		
		CASIA-WebFace	LFW	IJB-C
NbNet	19.2M	0.94	0.88	0.78
Proposed Method	5120×20	0.90	**0.91**	0.76
	1280×20	0.88	0.90	0.70
	320×20	0.75	0.73	0.54
	80×20	0.55	0.49	0.35
	40×20	0.34	0.27	0.22
	10×20	0.20	0.17	0.06

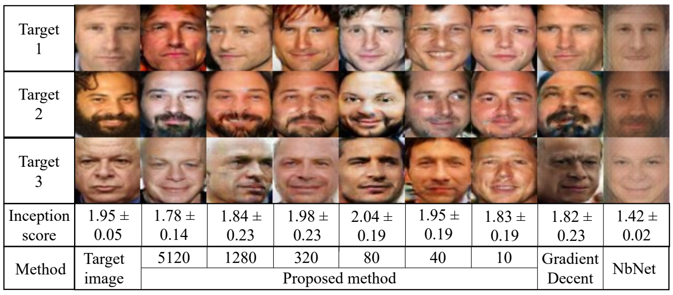
Rate (FAR), which indicates the probability that the comparison score between the reconstructed image and the corresponding evaluation template exceeds a threshold when it is presented to T. Note that the threshold was set to the value at which the target model's accuracy was the highest for the enrolled images in the template database.

4.2 Experimental results

Impact of the size of attack template. To evaluate the impact of the size of attack template m, we set the images per identity to 20 and varied the number of identity size from 10 to 5120. Therefore, m ranged from 10×20 to 5120×20. Table. 1 compares the our method against NbNet. Attack template size m for NbNet can be considered the same as its training dataset since the target feature extractor was assumed to be compromised in NbNet. We can see a positive correlation between the size of the available attack template and the SAR. Our method is comparable to NbNet in the perspective of the impersonation attack accuracy with a limited size of the attack template. Our method achieved an SAR of 88% with only 1280 identities, which was only 6% less than that of NbNet, given that the size of available attack templates was more than 720 times smaller for the `CASIA-WebFace`. The reconstruction attack with 320 identities is still effective, with a 75% chance of deceiving the target system.

Impact of the evaluation template. The environmental factor of images registered in the target system can be varied in a real world setting. Therefore, we also evaluated two more different datasets, `LFW` and `IJB-C`, while attack templates are calculated by `CASIA-WebFace`. Table. 1 also shows that our proposed method consistently performs well for datasets different from the training dataset. Notably, the attack for `LFW` with attack templates of 5120 and 1280 identities outperforms over NbNet by 3% and 2%, respectively. The relatively low SAR for `IJB-C` is presumably due to the difficulty of the recognition task of the dataset taking NbNet's SAR of 78% into account.

Reconstructing template. Fig. 3 shows the Inception Score and examples of reconstructed images of each method for the `LFW`. Our proposed method outperforms NbNet in terms of image quality regardless of the size of available attack templates m by virtue of exploiting a GAN independent of the compromised templates or target model. Moreover, the Inception Score of the proposed method was roughly equal to the original images of the `LFW` dataset.



Inception score	1.95 ± 0.05	1.78 ± 0.14	1.84 ± 0.23	1.98 ± 0.23	2.04 ± 0.19	1.95 ± 0.19	1.83 ± 0.19	1.82 ± 0.23	1.42 ± 0.02
Method	Target image	5120	1280	320	80	40	10	Gradient Decent	NbNet
		Proposed method							

Fig. 3: Reconstructed images on `LFW`. The first column shows the original target image. The second to seventh columns are the images reconstructed by our proposed method, with 5120, 1280, 320, 80, 40, and 10 identities, respectively. The eighth column is the images reconstructed by NbNet. The fourth row indicates the average Inception Score of each method with variance for the reconstructed 300 images.

5 Conclusion

In this study, we explored the possibility of a novel *black-box* attack, called a *model-free* template reconstruction attack, without the availability of a target feature extractor. We trained a feature converter to transform a biometric template from which we cannot reconstruct a face image into a biometric feature in a space associated with a model at the adversary's disposal. In an experimental trial, we have evaluated the impersonation accuracy and image quality of the reconstructed images compared with those of images reconstructed by the state-of-the-art *black-box* method. The experimental results showed that it is feasible to reconstruct a high quality image with a limited amount of pre-enrolled data without a dramatic decrease in attack accuracy.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP18H03239, 21H03442.

References

[ACB17] Arjovsky, M.; Chintala, S.; Bottou, L.: Wasserstein Generative Adversarial Networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 214–223, 2017.

[Ad03] Adler, A.: 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. volume 2, pp. 1163–1166 vol.2, 2003.

[Ad04] Adler, A.: Images can be regenerated from quantized biometric match score data. In: Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No.04CH37513). volume 1, pp. 469–472 Vol.1, 2004.

[De19] Deng, J; Guo, J; Xue, N; Zafeiriou, S: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2019.

[Do21] Dong, X.; Jin, Z.; Guo, Z.; Jin T., Andrew B.: Towards Generating High Definition Face Images from Deep Templates. In: 2021 International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–11, 2021.

[Du20] Duong, C. N.; Truong, T.; Luu, K.; Quach, K. G.; Bui, H.; Roy, K.: Vec2face: Unveil human faces from their blackbox features in face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6132–6141, 2020.

[FJ15] Fredrikson, M.; Jha, S.and Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. CCS '15. Association for Computing Machinery, pp. 1322–1333, October 2015.

[Hu12] Huang, G. B.; Mattar, M.; Lee, H.; Miller, E. L.: Learning to Align from Scratch. In: Advances in Neural Information Processing Systems (NIPS), 2012. 2012.

[KB15] Kingma, Diederik P; Ba, Jimmy: Adam: A Method for Stochastic Optimization. In: The International Conference on Learning Representations (ICLR). 2015.

[Li15] Liu, Z.; Luo, P.; Wang, X.; Tang, X.: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV). December 2015.

[Ma18] Maze, B.; Adams, J. et al.: IARPA janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). IEEE, pp. 158–165, 2018.

[Ma19] Mai, G; Cao, K; Yuen, P.C; Jain, A.K: On the Reconstruction of Face Images from Deep Face Templates. volume 41, pp. 1188–1202, May 2019.

[PVZ15] Parkhi, O. M.; Vedaldi, A; Zisserman, A: Deep Face Recognition. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, pp. 41.1–41.12, September 2015.

[Sa16] Salimans, T.; Goodfellow, I. et al.: Improved Techniques for Training GANs. In: Advances in Neural Information Processing Systems (NIPS). volume 29, 2016.

[So02] Soutar, C.: , Biometric system security. `http://www.bioscrypt.com/assets/security_soutar.pdf`, 2002.

[UJ04] Uludag, U.; Jain, A. K.: Attacks on biometric systems: a case study in fingerprints. In: Security, Steganography, and Watermarking of Multimedia Contents VI. volume 5306. SPIE, pp. 622–633, June 2004.

[Yi14] Yi, D.; Lei, Z.; Liao, S.; Li, S. Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923, 2014.

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, 23(10):1499–1503, 2016.

[Zh20] Zhang, Y.; Jia, R.; Pei, H.; Wang, W.; Li, B.; Song, D.: The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 2020.

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