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

DeDiM: De-identification using a diffusion model

Hidetsugu Uchida[1], Narishige Abe[2], Shigefumi Yamada[3]

Abstract: As a countermeasure against malicious authentication in a face recognition system using a face image obtained from SNS or the like, de-identification methods based on adversarial example have been studied. However, since adversarial example directly uses the gradient information of a face recognition model, it is highly dependent on the model, and a de-identification effect and image quality are difficult to achieve for an unknown recognition model. In this study, we propose a novel de-identification method based on a diffusion model, which has high generalizability to an unknown recognition model by applying minute changes to face shapes. Experiments using LFW showed that the proposed method has a higher de-identification effect for unknown models and better image quality than a conventional method using adversarial example.

Keywords: Face recognition, de-identification, diffusion model.

1 Introduction

Face is a famous biometric modality that enables identifying an individual from a face image, and it has rapidly spread because of the drastic accuracy improvements brought by deep learning technologies [De19]. However, since face images are easily available to a third person through SNS, video conferences, or the like, malicious authentication using face images and invasion of privacy have become significant issues [Sh21]. A common countermeasure to malicious authentication is to provide a face authentication system with presentation attack detection (PAD)[Mi20]. PADs are techniques for judging whether the input to the system is a bonafide sample. Because PADs are built into a face authentication system, they are not effective when a malicious user can select arbitrary authentication systems. It is also difficult, in principle, to handle injection attacks that directly input face image data to a system without going through the camera[Ca22]. Therefore, as a countermeasure for these realistic and advanced scenarios of malicious authentication, image processing techniques that make a face image itself difficult to identify by a face authentication system, that is, de-identification are under a growing demand.

Importantly, in de-identification, the visual information of an original image must not be lost because of the de-identification process, considering the original purpose of the face image such as SNS or video conferencing. De-identification methods using adversarial example have drawn attention as techniques that should achieve de-identification effects and visual information preservation[Ya21]. However, an adversarial sample has a high

[1] Fujitsu Ltd., 4-1-1 Kamikotanaka, Nakahara-ku, Kawasaki-shi, Kanagawa, Japan, u.hidetsugu@fujitsu.com
[2] Fujitsu Ltd., 4-1-1 Kamikotanaka, Nakahara-ku, Kawasaki-shi, Kanagawa, Japan, abe.narishige@fujitsu.com
[3] Fujitsu Ltd., 4-1-1 Kamikotanaka, Nakahara-ku, Kawasaki-shi, Kanagawa, Japan, yamada.shige@fujitsu.com

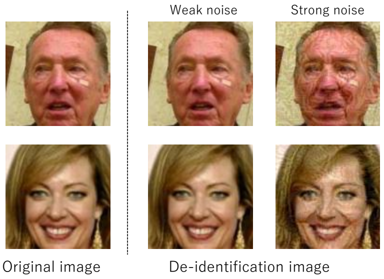

Fig. 1: Examples of TIP-IM.

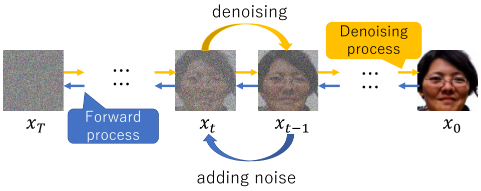

Fig. 2: DDPM flow

model dependency because the gradient information of a face recognition model is directly applied to an image, and the de-identification effect on unknown models (models other than the model considered in image processing) tends to be reduced.

Therefore, in this study, we propose a method of de-identification by making minute changes to face shapes. Since face shapes are a feature that various face authentication systems are commonly considered to refer to, by imparting a de-identification effect to the shapes, a high generalization is expected. In this study, the denoising diffusion probabilistic model (DDPM)[HJA20], which is a diffusion model, is used to generate de-identification images. DDPMs are suitable techniques as a generator in de-identification because they can generate a high-quality and high-variety image while appropriately retaining the original image information. The contributions of our work are as follows:

- We propose a de-identification method using a diffusion model(DeDiM). Unlike adversarial example-based methods, because the proposed method gives a change having a de-identification effect to the shapes of face parts, high generalization to unknown models can be expected.
- Comparative experiments with a conventional method using adversarial example and DeDiM show that DeDiM is superior in retrieval and verification when an unknown model is used.

2 Related works

Adversarial example [GSS14] uses generated noise to induce misclassification of a recognition model. The noise is generated based on the gradient information of the recognition model and is superimposed on an input (e.g. image). In a previous study of de-identification, the targeted identity-protection iterative method (TIP-IM)[Ya21], used the gradient information of a loss function that minimizes the similarity of the feature vector, that is, outputs of a face recognition model, between the input image and the processed image and maximizes the similarity of that between the processed image and a target image whose identity differs from that of the input image to generate noise. The noise is constrained not only by the similarity of the feature vector (i.e. the identity) but also by the naturalness of the processed image considered by intensity of the noise and the maximum mean discrepancy (MMD)[Gr12] between input images and processed images. By controlling this constraint on naturalness, the amount of visual information of the input image maintained in the processed image can be controlled.

However, de-identification using adversarial example tends to depend on a face recognition model because the gradient information of the model is used as noise. For unknown models, the effect of de-identification tends to weaken. A TIP-IM study has evaluated the effect on unknown models, and it has been reported that although the effect is improved compared with other conventional methods, the effect appears to be reduced compared with the known models. To have a high effect on an unknown model, the intensity of the noise must be increased. Fig. 1 shows TIP-IMs examples: weak noise images and strong noise images. As shown in Fig. 1, when the intensity of the noise is strong, stripe patterns are generated and degrade image quality. Because TIP-IM noise has no constraints on visual behavior except the constraint on the naturalness defined by the MMD, the change of pixel information caused when the noise intensity is increased (when the de-identification effect dominates the naturalness) is not guaranteed to become natural in the face image. That is, this problem is caused by insufficient consideration of the structure of the face image in the de-identification processing.

Deep generative models such as GAN (generative adversarial network)[Ka21] and VAE (variational auto-encoder)[ROV19] can generate high-quality face images. Methods for flexibly controlling the individuality, expression, and pose of a face image using these generative models have been proposed, and various applications such as the anonymization of a face image have been studied[LL19]. Diffusion models can generate face images of comparable quality to GAN or VAE-based models. In diffusion models, an image is sampled by repeatedly applying a denoising process to noisy image (see Fig.2). This sampling process can start with not only a complete noise image but also a given image (the original image) with added noise. In this case, as the noise is smaller, an image similar to the original image is sampled. This property is advantageous from the viewpoint of preserving the visual information of the input image in de-identification. Furthermore, methods for controlling a class of sampled images using a discriminative model have been studied (classifier guidance sampling[DN21]). Using guidance sampling, the images belonging to an intended class can be sampled from a diffusion model that can generate many classes of images. In this study, we use this guidance sampling to control the identity of a face image.

3 Method

DeDiM uses two types of deep learning models. One is a generative model for processing face images. The other is a face recognition model for controlling the individuality of the processing result.

As generative models, DDPMs are used. Then, we briefly review the formulation of DDPMs. DDPMs sample an image using a process in which noise is repeatedly superimposed on a

Algorithm 1 Sampling process in DDPM

```
input $\theta$
$\boldsymbol{x_T} \sim N(\mathbf{0}, \boldsymbol{I})$
for $t = T, \ldots, 1$ do
  $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
  $\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x_t} - \frac{\beta_t}{\sqrt{1-\hat{\alpha}_t}}\varepsilon_\theta(\boldsymbol{x_t}, t)) + \sigma_t \boldsymbol{z}$
end for
return $\boldsymbol{x_0}$
```

Algorithm 2 Sampling process in DeDiM

```
input $\boldsymbol{x}_{org}$ $\theta$, $\phi$, $t_0$, m ,$s$, $L$, $h$
$\boldsymbol{x}_0^{(1)} \leftarrow \boldsymbol{x}_{org}$
for $l = 1, \ldots, L$ do
  $\boldsymbol{x}_{t_0} \sim N(\boldsymbol{x}_{t_0}; \sqrt{\hat{\alpha}_t}\boldsymbol{x}_0^{(l)}, (1-\alpha_t)I)$
  for $t = t_0, \ldots, 1$ do
    $\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = 0$
    $\boldsymbol{x}_{t-1} = \mu_\theta(\boldsymbol{x_t}, t)$
                $+ s\sigma_t \nabla_{\boldsymbol{x_t}} \log p(\tilde{y_0}|\boldsymbol{x_t}) + \sigma_t \boldsymbol{z}$
  end for
  $\boldsymbol{x}_0' = m \circ \boldsymbol{x}_0 + (1 - \boldsymbol{m}) \circ \boldsymbol{x}_{org}$
  $\boldsymbol{x}_0^{(l+1)} \leftarrow \boldsymbol{x}_0'$ if $p(y_0|\boldsymbol{x}_0') < p(y_0|\boldsymbol{x}_0^{(l)})$,
          else $\boldsymbol{x}_0^{(l+1)} \leftarrow \boldsymbol{x}_0^{(l)}$
  break if $p(y_0|\boldsymbol{x}_0^{(l+1)}) < h$
end for
return $\boldsymbol{x_0}^{(l+1)}$
```

clean image, resulting in a complete noise image. This "forward process" is defined as

$$q(\boldsymbol{x_1}, \ldots, \boldsymbol{x_T}|\boldsymbol{x_0}) := \prod_{t=1}^{T} q(\boldsymbol{x_t}|\boldsymbol{x}_{t-1}), \tag{1}$$

$$q(\boldsymbol{x_t}|\boldsymbol{x}_{t-1}) := N(\boldsymbol{x_t}; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}). \tag{2}$$

Here, $\boldsymbol{x_0}$ is a clean image sampled from a given data distribution $q(\boldsymbol{x_0})$. $\boldsymbol{x_t}$ is an image at time step t, and T is the total step size of the forward process (noise is added T times in total). $\beta_t \in (0,1)$ represents a variance of noise. When T is sufficiently large and β_t is set to an appropriate value, $\boldsymbol{x_T}$ follows $N(\mathbf{0}, \boldsymbol{I})$. Once a "denoising process" , $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x_t})$ is modeled, the clean image $\boldsymbol{x}_0$ can be sampled by repeatedly applying the denoising process to the noise image $\boldsymbol{x}_T$. The denoising process can be defined as

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x_t}) := N(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x_t}, t), \sigma_t^2 \boldsymbol{I}). \tag{3}$$

Here, $\mu_\theta(\boldsymbol{x_t}, t)$ is represented as $\frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x_t} - \frac{\beta_t}{\sqrt{1-\hat{\alpha}_t}}\varepsilon_\theta(\boldsymbol{x_t}, t))$. α_t and $\hat{\alpha}_t$ are $1-\beta_t$ and $\prod_{s=1}^{T}\alpha_s$ respectively, and σ_t^2 is $\frac{1-\hat{\alpha_{t-1}}}{\hat{\alpha}_t}\beta_t$. $\varepsilon_\theta(\boldsymbol{x_t}, t))$ is modeled with a neural network (denoising network). The sampling process for clean image is shown in Algorithm 1. As a derivation of the sampling process of DDPM, a classifier guidance sampling has been studied[DN21]. The sampling uses a classifier to sample a image in an intended class that the classifier defines. Let $p_\phi(y|\boldsymbol{x_t})$ be a posterior of class y. In th guided sampling, $\mu_\theta(\boldsymbol{x_t}, t)$ in Eq.3 is replaced with $\mu_\theta(\boldsymbol{x_t}, t) + s\sigma_t^2 \nabla_{\boldsymbol{x_t}} \log p_\phi(y|\boldsymbol{x_t})$. Here, s is the gradient scale that controls the forcing of the guide.

In DeDiM, the classifier guidance sampling is applied with de-identification. When a face image $\boldsymbol{x_0}$ having identity y is given, the goal of de-identification is to enable predicting y from the image. In other words, when the image is classified other than y, de-identification is achieved. Thus, if we can know that probability, the de-identification image can be sampled by using the guided sampling with that probability. However, becase it is generally difficult to construct a classifier corresponding to an arbitrary identity y, a face recognition model is used instead of a classifier.

A face recognition model is a feature extractor, that is represented as $\boldsymbol{v} = f_{\phi}(\boldsymbol{x})$. Here $\boldsymbol{v}$ is a feature vector of $\boldsymbol{x}$. Given an original image $\boldsymbol{x_0}$ for de-identification, the probability that the sampled image $\boldsymbol{x_t}$ is classified into the same class as the original image can be approximated as

$$p(y_0|\boldsymbol{x_t}) \approx (\cos(\boldsymbol{v_t}, \boldsymbol{v_0}) + 1)/2. \tag{4}$$

Here, y_0 is identity of the original image and $\cos(\cdot,\cdot)$ is the cosine similarity. The probability that $\boldsymbol{x_t}$ is classified other than y_0 is represented as $p(\tilde{y_0}|\boldsymbol{x_t}) = 1 - p(y_0|\boldsymbol{x_t})$. Using $p(\tilde{y_0}|\boldsymbol{x_t})$ as the classifier in the sampling, we can sample a face image that has an identity other than y_0. Note that, if the sampling is started from time step T, the sampled image has a visually different identity from that of the original image. Therefore, in DeDiM, to preserve the visual information of the original image, the sampling process is started from a time step much smaller than T. This approach is because how much visual information from the original image remains in the sampled image depends on the time step in which the sampling process begins. That is, if a time step close to T is set as the starting point, the information of the original image is barely retained, and conversely, if it is close to 0, much information about the original image is retained, and only the local fine structure changes. The de-identification effect of the proposed method is embedded in the fine structure of the original image. The sampling process of the proposed method is shown in Algorithm 2. In Algorithm 2, $\boldsymbol{x_{org}}$ is the original face image and t_0 is the starting time step. The $\boldsymbol{m}$ is an image mask that has the same size as the original image, and the values around the facial landmarks (eyes, eyebrows, nose, and mouth) of the original image are 1, and the other pixels are 0. By combining the original image and the sampled image using the image mask, the pixel information of the original image can be used as it is. In addition, to enhance the de-identification effect, the sampling process is repeatedly applied to the combined image until the probability that the combined image is classified to the original identity falls below the threshold h or reaches a predetermined number of iterations L is reached.

4 Experiments

To evaluate the performance of DeDiM, experiments on the retrieval accuracy and verification accuracy for de-identification images were conducted. The retrieval experiments were designed referring to previous research[Ya21]. In the retrieval experiment, a set of face images including a plurality of IDs was prepared as the gallery set, and the IDs in the

	retrieval				verification	
	Rank-1	Rank-1	PSNR	MSSIM	FNMR(WB)	FNMR(BB)
method	(WB)	(BB)			(FMR 0.1%)	(FMR 0.1%)
orignal image	0.818	0.892	-	1.00	0.17	0.10
TIP-IM[Ya21]	0.000	0.254	29.3	0.78	1.00	0.77
DeDiM	0.008	0.124	28.1	0.91	0.99	0.85

Tab. 1: results of the retrieval and verification experiments

gallery were retrieved from the face images given as queries. LFW (Labeled Face in the Wild)[Hu07] was used as the experimental data. From LFW, 500 images having different IDs were selected as queries and 1000 images that contain 500 images having the same IDs as queries and 500 images having IDs other than queries were selected as the gallery set. A face recognition model was employed to extract to a feature vector from a face image, and a similarity of vectors between a query and a gallery image was used for retrieval. Then, the de-identification effect was evaluated from the ratio that the ID corresponding to the query is included in the upper rank of a retrieval result. On the other hand, in the verification experiments, FMR(False Match Rate) and FNMR(False Non-Match Rate) were calculated from genuine scores and imposter scores between the queries and the gallery images. All face images used in the experiment were pre-processed by MTCNN (multi-task cascaded convolutional neural network)[Zh16] to detect face areas and formed into alignment images of 112 $\times$112 pixels.

In the set up for DeDiM, the DDPM referred to previous studies[HJA20]. The denoising network was learned with the Flickr-Faces-HQ Dataset[Fl22]. T was set to 1000, and a cosine schedule[ND21] was used for β. As a face recognition model for the guided sampling, mobilefacenet[Ch18] was used. The gradient scale s was 250, the starting point t_0 was 100, the maximum number of loop L is 100 and the threshold h was 0.2. The image mask m was made using landmarks extracted using MTCNN, where the pixels within 10 pixels from each of 51 (Eyes 12 points, eyebrows 10 points, nose 9 points, and mouth 20 points) landmarks were defined as 1. As a comparison method, we evaluated TIP-IM, which is the state-of-the-art of de-identification. In TIP-IM, to generate noise, mobilefacenet was also used as the proposed method[3].

Tab.1 shows the results of the retrieval and verification experiments. Rank-1(WB) is the rank-1 accuracy in the white box condition such that the feature vector used in the retrieval and de-identification process is the same, that is, mobilefacenet was used for the de-identification process and retrieval. On the other hand, Rank-1(BB) is the rank-1 accuracy in the black box condition such that the feature vectors used in the retrieval and the de-identification process are different. SE-Net50[Hu20] was used as the face feature for retrieval in the black box condition.

Each value of the retrieval in Tab.1 is an average value of 500 query images. A lower Rank-1 accuracy means that the effect of de-identification is more effective, while a higher PSNR

[3] As a target image to be used for noise generation, face images having IDs not included in queries and registered images were randomly selected from LFW.

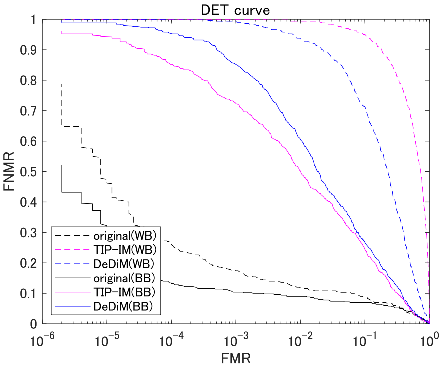

Fig. 3: DET curve

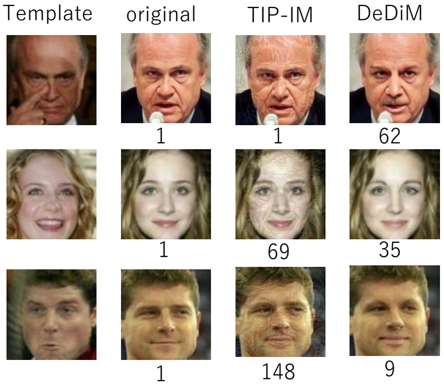

Fig. 4: Examples of de-identification.

(peak signal-to-noise ratio) and MSSIM (mean structural similarity)[Wa04] indicate that the visual information of the original image is well maintained. The results of TIP-IM showed that the Rank-1 accuracy in WB was reduced to 0 (completely undifferentiated), while in BB it was significantly higher than that in WB, although lower than the original images[4]. This results indicates that TIP-IM does not have sufficient generalization. On the other hand, in DeDiM, the Rank-1 accuracy in WB is sufficiently low, and in BB, the increase in the Rank-1 accuracy is suppressed compared with TIP-IM. In addition, PSNR and MSSIM in DeDiM are equal to or higher than those of TIP-IM. These results clarified that DeDiM can exhibit a high de-identification effect for an unknown model while maintaining visual information. The results of verification experiments, FNMRs when FMR is 0.1%, are shown in Tab.1 and the DET curve is shown in Fig3. These results also support that DeDiM has higher generalization performance for unknown models than TIP-IM. Fig.4 shows examples of the de-identification images of DeDiM and TIP-IM. The numbers below the images are the ranks in the retrieval experiment (BB). TIP-IM images show noticeable changes due to noise, resulting in unnatural images. On the other hand, in DeDiM, the information of the original image is preserved and the face image is natural.

5 Conclusion

This paper proposed a novel de-identification method using a diffusion model, DeDiM. Experiments clarified that the de-discrimination effect of DeDiM showed higher generalization performance for an unknown model than that of a conventional method based on adversarial example in retrieval and verification. In addition, good results were obtained for preserving of visual information. However, because the evaluation used objective evaluation methods, the evaluation for preserving visual information by subjective evaluation is future work.

[4] Since the data set used in this experiment contains several label and detection errors, the results in the original images are worse than the benchmark result in general LFW.

References

[Ca22] Carta, K.; Barral, C.; Mrabet, N. El; Mouille, S.: Video injection attacks on remote digital identity verification solution using face recognition. In: 13th International Multi-Conference on Complexity, Informatics and Cybernetics. 2022.

[Ch18] Chen, S.; Liu, Y.; Gao, X.; Han, Z.: MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In: Chinese Conference on Biometric Recognition. 2018.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Conference on Computer Vision and Pattern Recognitio. 2019.

[DN21] Dhariwal, P.; Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. arXiv preprint arXiv:2105.05233, 2021.

[Fl22] Flickr-Faces-HQ Dataset, https: //github.com/NVlabs/ffhq-dataset.

[Gr12] Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; Smola, A.: A kernel two-sample test. Journal of Machine Learning Research, pp. 723–773, 2012.

[GSS14] Goodfellow, I.; Shlens, J.; Szegedy, C.: Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572, 2014.

[HJA20] Ho, J.; Jain, A.; Abbeel, P.: Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2006.11239, 2020.

[Hu07] Huang, G.; Mattar, M.; Berg, T.; Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: International Conference on Computer Vision. 2007.

[Hu20] Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E.: Improved Denoising Diffusion Probabilistic Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 2011–2023, 2020.

[Ka21] Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T.: Alias-Free Generative Adversarial Networks. In: 35th Conference on Neural Information Processing Systems. 2021.

[LL19] Li, T.; Lin, L.: Natural Face De-Identification with Measurable Privacy. In: Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[Mi20] Ming, Z.; Visani, M.; Luqman, M.; Burie, J.: Survey on Anti-Spoofing Methods for Facial Recognition with RGB Cameras of Generic Consumer Devices. Journal of Imaging, 2020.

[ND21] Nichol, A.; Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In: the 38 th International Conference on Machine Learning. 2021.

[ROV19] Razavi, A.; Oord, A.; Vinyals, O.: Generating Diverse High-Fidelity Images with VQ-VAE-2. arXiv preprint arXiv:1906.00446, 2019.

[Sh21] Shopon, M.; Tumpa, S.; Bhatia, Y.; Kumar, K.; Gavrilova, M.: Biometric Systems De-Identification: Current Advancements and Future Directions. Journal of Cybersecurity and Privacy, pp. 470–495, 2021.

[Wa04] Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, pp. 600–612, 2004.

[Ya21] Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; Xue, H.: Towards Face Encryption by Generating Adversarial Identity Masks. In: International Conference on Computer Vision. IEEE, 2021.

[Zh16] Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y.: Joint Face Detection and Alignment Using Multi-task Cascaded Convolutional Networks,. IEEE Signal Processing Letters, pp. 1499–1503, 2016.